AN ENGINEERS' MANUAL OF STATISTICAL METHODS

BY

LESLIE E. SIMON

Major Ordnance Department, U. S. Army
Assistant Director, The Ballistic Research Laboratory

NEW YORK
JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED
1941

PRINTED IN U. S. A.

PRESS OF
BRAUNWORTH & CO., INC.
BUILDERS OF BOOKS
BRIDGEPORT, CONN.

PREFACE

There is no useful mathematical weapon
which an engineer may not learn to use.
—Lord Kelvin

This book does not purport to be a formal textbook on probability, statistics, or logic.[1] It is a summary of certain working parts of these sciences, designed for the practical man whose interest is centered in his industrial or engineering work. It endeavors to place in his hands a few practical working tools which, without any serious alteration in his working principles, will enable him to do what he wants to do more quickly, surely, and economically.

The presentation, therefore, is as brief as is compatible with clarity. Controversial issues and statistical methods of lesser practical importance to industrial and engineering work are neglected except so far as the discussion of them contributes to development of important methods. Proofs, elucidating explanations, and methods too important to be omitted altogether are minimized or covered in appendices.

The principles relating to inspection methods and statistical inference are predicated on recognized standard procedures or on research conducted at the Ballistic Research Laboratory, Aberdeen Proving Ground, incident to various ordnance problems, but principally in connection with designing an inspection procedure for the Army for determining the quality of reserve stocks of ammunition which have suffered partial deterioration. The principles relating to process inspection and control of manufactured products are predicated on like grounds and on experience at a Government manufacturing arsenal. Illustrative examples from actual experience are altered so as to convey no information of a military or engineering nature, but not in a way to prejudice their illustrative value.

It is a pleasure to acknowledge my indebtedness to Dr. Walter A. Shewhart of the Bell Telephone Laboratories, Dr. W. Edwards Deming of the Department of Agriculture Graduate School, and Mr. A. Adele-

[1] Some relevant textbooks on these subjects are listed under Literature Cited.

man of the Technical Staff, Office of the Chief of Ordnance, without whose unselfish counsel, advice, and criticism this work would not be possible. I am also indebted to various members of the staff of the Ballistic Research Laboratory, and the officers stationed at Aberdeen Proving Ground, for many helpful suggestions.

This book is written primarily for use in the Ordnance School, Aberdeen Proving Ground, for instructing Ordnance officers in the minimum of statistical techniques necessary to their functions as engineers, supervisors, and executives in technical and industrial work. It is believed, however, that it is of especial value to professional men in like fields to whom statistics is merely one of a number of convenient tools for accomplishing their professional ends.

LESLIE E. SIMON

THE BALLISTIC RESEARCH LABORATORY
ABERDEEN PROVING GROUND, MARYLAND
January, 1941

CONTENTS

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER V

CHAPTER VI

CHAPTER VII

CHAPTER VIII

CHAPTER IX

CHAPTER X

CHAPTER XI

CHAPTER XII

APPENDIX A

APPENDIX B

APPENDIX C

APPENDIX D

AN ENGINEERS' MANUAL OF STATISTICAL METHODS

CHAPTER I

INTRODUCTION
MAKING SENSE OUT OF FIGURES

Egad, I think the interpreter is the hardest
to be understood of the two!
—RICHARD BRINSLEY SHERIDAN: *The Rivals*

The advance of statistical methods. In the last few years the increasing use of statistical methods has been too striking to have escaped one's attention, whatever his business or profession. A little over a decade ago their use was confined almost exclusively to biometrics and the social sciences. At the present time, however, their application to engineering and manufacturing has been so successfully developed that to ignore them is sheer folly.

The development of any new product or new idea requires an initial outlay which yields no immediate return. It is not surprising, therefore, that the initial development of statistical methods for industrial and engineering use was confined almost exclusively to a few corporations capable of maintaining large research staffs. The pioneers in this field were prompt in employing the product of their research with great economic advantage to themselves. Their methods were investigated and their lead followed by many others of the larger corporations so that, at the time of the present writing, a list of the organizations using statistical methods reads like a *Who's Who* in American industry.[1]

The fruit of this research and the guidance of its subsequent use which has proved its value are now available to all. However, the literature on the subject is written largely by and for statisticians. The mode of thought and terminology are largely unfamiliar to the uninitiated, and the practical man is likely to wonder what it is all

[1] See *Proceedings of the Industrial Statistics Conference*, Pitman Publishing Corp., New York, 1939.

about and why he as an executive, engineer, or industrialist should adopt these new methods when he is probably doing very well as he is. Stripped of their verbiage, statistical methods become very easy, and their application is still easier. It is believed that a simple illustration will show the plain, logical common sense behind statistical methods, demonstrate their mode of operation, and at least indicate some important ways in which they can be of service to the practical man.

An ordinary sampling problem. Everyone is concerned with drawing inferences from samples. Sampling is the heart and soul of almost all specifications; and almost everyone is concerned directly or indirectly with specifications. Sampling may be classified in two ways: sampling by attributes and sampling by variables. The former relates to the classification of articles in one of two ways; e.g., those which pass the go, not-go gauge and those which do not pass the go, not-go gauge; those which function as intended and those which do not function as intended; in short, those which possess any assigned characteristic and those which do not possess that characteristic. On the other hand, sampling by variables refers to classification made on a continuous scale such as the length of a piece part, the burning life of an electric lamp, the blowing time of a fuse, etc. Since sampling by variables is somewhat more complicated than sampling by attributes, let us first consider the simple case where 10 articles are selected at random from a large lot, batch, or consignment of articles and, upon inspection or test, 9 prove good and 1 defective. The sample might be larger or smaller, and the number of defectives greater or zero; 10 and 1 are selected merely for the purpose of convenient illustration and because many acceptance specifications use these actual figures, i.e., require that a sample of 10 shall be selected at random and not more than 1 shall fail. This problem may be regarded as a sample taken from one's production line for the purpose of estimating whether or not a current batch of product is of sufficient quality to warrant its being sent out to service; it may be a sample from a consignment of goods that is offered for acceptance; it may represent the test of a few pilot models of a newly developed article or apparatus; or it may be a specification that one's product must meet, and naturally one wonders how good he must make his product in order that it will have a fair chance of passing the specification. The important question is, what does such a type of sample or requirement mean?

Statistics has been defined as the art of making common sense out of figures. Therefore, it is proposed to examine this requirement in the light of ordinary unaided judgment, and in the light of statistical analysis. This critical examination will reveal two rather startling phenomena. The first might be called "things are not what they seem," for the discussion will show that small samples taken from a moderately defective lot of articles are better than the lot more frequently than they are poorer than the lot. This may sound strange, so let me repeat it in figures. For example, if a lot of articles consists of 9000 good articles and 1000 defective articles, and if this lot is repeatedly sampled by random samples of 10, samples which have a higher percentage of good articles than the lot will occur considerably more frequently than samples which have a lower percentage of good articles than the parent lot. The discussion of the second phenomenon will show that specifications of the above type do not specify anything in particular about the quality of the products accepted thereunder, and that the accepted quality is practically that which happens to be offered for acceptance. The discussion will also show that, from an economic point of view, the manufacturer need not make the best possible product to meet such a specification, but instead may consciously set his level of quality so that the cost of manufacture plus the cost of relatively infrequent rejections is a minimum.

Casual judgment. In the light of ordinary unaided judgment one would be likely to infer that a sample which is 10% defective came from a lot which is 10% defective. Likewise, it might appear that a specification of the above type would insure the acceptance only of lots of articles which contain at least 90% good articles. However, a moment's reflection shows that this premise may be untrue, for the percentage of good articles in a sample (or sample fraction effective) is not necessarily the same as the percentage of good articles in the lot (or lot fraction effective). The influence of chance in the selection of samples precludes this very simplifying situation. Consequently, some lots may be accepted on a basis of sample fraction effective equal to or greater than 0.9, when the lot fraction effective is less than 0.9. Hence, these rather apparent estimates are faulty. Suppose, then, that the casual estimate is made in a more conservative manner, and let one say that lots of 0.9 fraction effective will be accepted only 50% of the times that they are presented. This estimate, of course, implies that, in lots of fraction effective 0.9, the sample fraction

effective will be greater than the lot fraction effective just as frequently as it will be less than the lot fraction effective. Such a relationship cannot be accepted without proof; and it therefore appears proper that one should inquire into the relationship between large lot and small sample before attempting to draw an inference from the observation of 1 defective in a sample of 10 or to place an interpretation on the requirement that a sample of 10 shall not contain more than 1 defective.

Relationship between small sample and large lot. Suppose then that a large lot of articles is fraction effective 0.9. If a random sample of 10 is drawn from this lot, one can figure by simple combinations the probability that a random sample of 10 will contain 10 good articles and 0 defective ones, 9 good and 1 defective articles, 8 good and 2 defective articles, · · · 0 good and 10 defective articles. The method is so simple and amusing that it is well worth giving.

If the fraction effective of the lot is 0.9, then 9/10 of the articles in the lot are good articles. If a single article is drawn at random, the probability that it will be a good article is obviously 0.9. If 2 successive articles are drawn, the probability that they will both be good articles is manifestly 0.9×0.9 or 0.81. Similarly, if 10 successive articles are drawn, the probability that all 10 will be good is $(0.9)^{10} = 0.35$. The probability of drawing 9 good articles and 1 defective article is scarcely so easy, but not difficult at that. Such a sample can be drawn by first drawing 1 defective article and then drawing 9 good articles. The probability of drawing 1 defective article on the first draw is 0.1. The probability of drawing 9 subsequent good articles is $(0.9)^9$. Thus the probability of 1 defective article followed by 9 good articles is $(0.1)(0.9)^9$ or 0.039. This same net sample can be obtained by first drawing 1 good article, then 1 defective article, and then 8 good articles. The probability of this event is $(0.9)(0.1)(0.9)^8$, or again 0.039. If this reasoning be continued for 2 good articles, 1 defective article, and 7 good articles, etc., one finds that there are 10 different ways of drawing a net sample of 9 good and 1 defective articles, and that the probability in each case is 0.039. Therefore the probability of drawing 9 good articles and 1 defective article in any one of the 10 ways (which one is immaterial) is $10(0.9)^9(0.1) = 0.39$. In like manner it can be shown that there are 45 ways of drawing 8 good articles and 2 defective articles, the probability of each of which is $(0.9)^8(0.1)^2$, or a total probability of $45(0.9)^8(0.1)^2 = 0.19$. The process of arranging these combinations soon becomes very laborious;

but it can be readily avoided, for one can see that the probability of drawing c defective articles in a sample of n is merely the number of combinations that can be made of n things taken c at a time, multiplied by the probability of drawing $(n - c)$ successive good articles, times the probability of drawing c successive defective articles. Now, by fortuitous circumstance, it happens that the first, second, third, etc., terms of the expansion of the binomial $(0.9 + 0.1)^n$ are precisely these numbers of combinations times these respective probabilities for $c = 0$, $c = 1$, $c = 2$, etc. Hence, the probability of drawing c defectives in a random sample of n articles is merely the $(c + 1)$th term of the binomial. If one happens to remember the binomial theorem he can write the expansion as:

$$(0.9)^n + \frac{n}{1}(0.9)^{n-1}(0.1) + \frac{n(n-1)}{1 \times 2}(0.9)^{n-2}(0.1)^2 \cdots$$

The general formula for the $(c + 1)$th term can be conveniently written as:

$$\frac{n!}{(n-c)!\,c!}(0.9)^{n-c}(0.1)^c,$$

where $n!$ is the product of all the natural numbers from 1 to n inclusive and $0!$ is by definition 1. Thus, if Q is the fraction defective of of the lot, and therefore $1 - Q = P$ is the fraction effective, then one can at once calculate the probability of drawing c defective articles in a sample of n from the simple expression [2]

$$\frac{n!}{(n-c)!\,c!}P^{n-c}Q^c,$$

which is easy when n is small but becomes increasingly difficult, even with the aid of tables, when n is large. Pursuant to this well-established procedure, Fig. 1·1 has been drawn to show the percentage of times that random samples of 10 will be fraction effective 0.0, 0.1, 0.2, · · · 1.0, when drawn at random from a large lot of fraction effective 0.9. The asymmetry of the curve is significant. It should be noted that a lot of fraction effective 0.9 will yield a sample which is perfect 35% of the time, a sample which is 0.9 fraction effective 39% of the time, and a sample which is of poorer fraction effective than the lot only 100% − 39% − 35% or 26% of the time. (These

[2] For a thorough discussion of the binomial, see *Probability and Its Engineering Uses*, T. C. Fry, D. Van Nostrand, New York, 1928.

exact figures are 0.3486784401, 0.387420489, and 0.2639010709.) Thus, it is obvious that samples of 10 taken from a large lot which is 10% defective are better than the lot considerably more frequently than they are poorer than the lot.

An additional line on Fig. 1·1 shows the distribution of samples of 10 from a lot which is fraction effective 0.5. Note that this curve is perfectly symmetrical; and there is no tendency for the sample to misrepresent the lot, since samples which are better than the lot occur just as frequently as samples which are poorer than the lot. Although rigorous proof would be tedious, a study of these curves

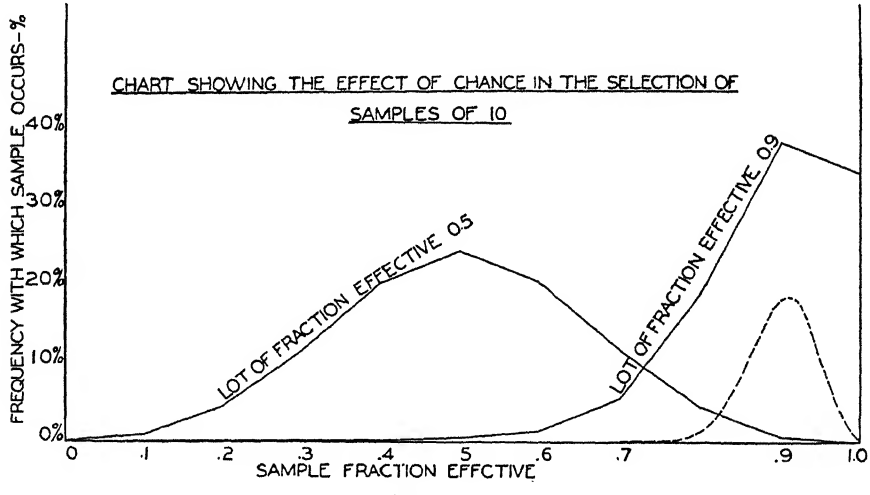


FIG. 1·1.

will show that, the nearer the lot fraction effective is to 0.5, the less skewed the curve, and the less the discrepancy between the frequency of samples which are better than the lot and the frequency of samples which are poorer than the lot. Similarly, the dotted line on Fig. 1·1 showing the distribution for samples of 50 from the lot of fraction effective 0.9 shows that the larger the sample size the less skewed the curve and the less the havoc played by chance in the selection of samples. That is to say, if the sample is large enough, or if the fraction defective of the lot is close enough to 0.5, this difficulty is overcome. However, these are seldom the conditions met in practice. Thus the statement that small samples taken from a moderately defective lot will be better than the lot more frequently than poorer than the lot is proved. Hence, having observed 1 defective

in a sample of 10, one should rather suspect that the lot contains more than 10% defectives.

A common-sense check on the relationship between sample and lot. Someone is likely to remark that the proof is only mathematical and does not make common sense. Let us see if it does. Suppose that the lot consisted of 10,000 articles, 9000 of which were good and 1000 of which were defective; and that 1000 samples of 10 were taken; i.e., the lot was sampled until exhausted. The average of the sample fractions effective would of necessity have to equal 0.9. Now note that the sample can be better than the lot only by being perfect: a margin of 1 unit. It can be poorer than the lot by containing 0, 1, 2 · · · 8 good articles, which is a margin of up to 9 units. For every sample containing only 7 good articles (2 below average), 2 perfect samples must be drawn to make up for it; for every sample containing only 6 good articles, 3 perfect samples must be drawn, etc. Thus, when the lot fraction effective is greater than 0.5, it stands to reason that samples which are better than the lot must predominate in frequency over samples which are poorer than the lot, because the occurrence of just 1 sample (in the present example) of 7 or less good articles would force this predomination. Statistics will always check with common sense, or there is something the matter with the statistics.

A physical demonstration of the relationship of sample to lot. Occasionally some proponent of the proof-of-the-pudding-is-in-the-eating attitude remains unconvinced by these arguments. To convince these people by means of gestures, one of the laboratory mechanics constructed the sampling machine shown in Plate I. At the beginning of a discourse, such as that just given, the hopper of the machine is loaded with 900 copper-plated steel balls and 100 chromium-plated steel balls (the latter being regarded as defectives). The adjustable mechanism on the front of the machine is set for sample size 10, and a spectator is invited to operate the machine. The steel balls are continually agitated during sampling. The samples of 10 flow into the slot shown on the face of the machine. The operator can tell at a glance whether the sample is better than the lot (perfect), the same as the lot (1 defective in 10), or poorer than the lot (more than 1 defective in 10). By rotating the sampling disk, samples which are better than the lot are caused to flow into the right-hand test tube; samples which are the same as the lot into the middle test tube; and samples which are poorer than the lot into the left-hand

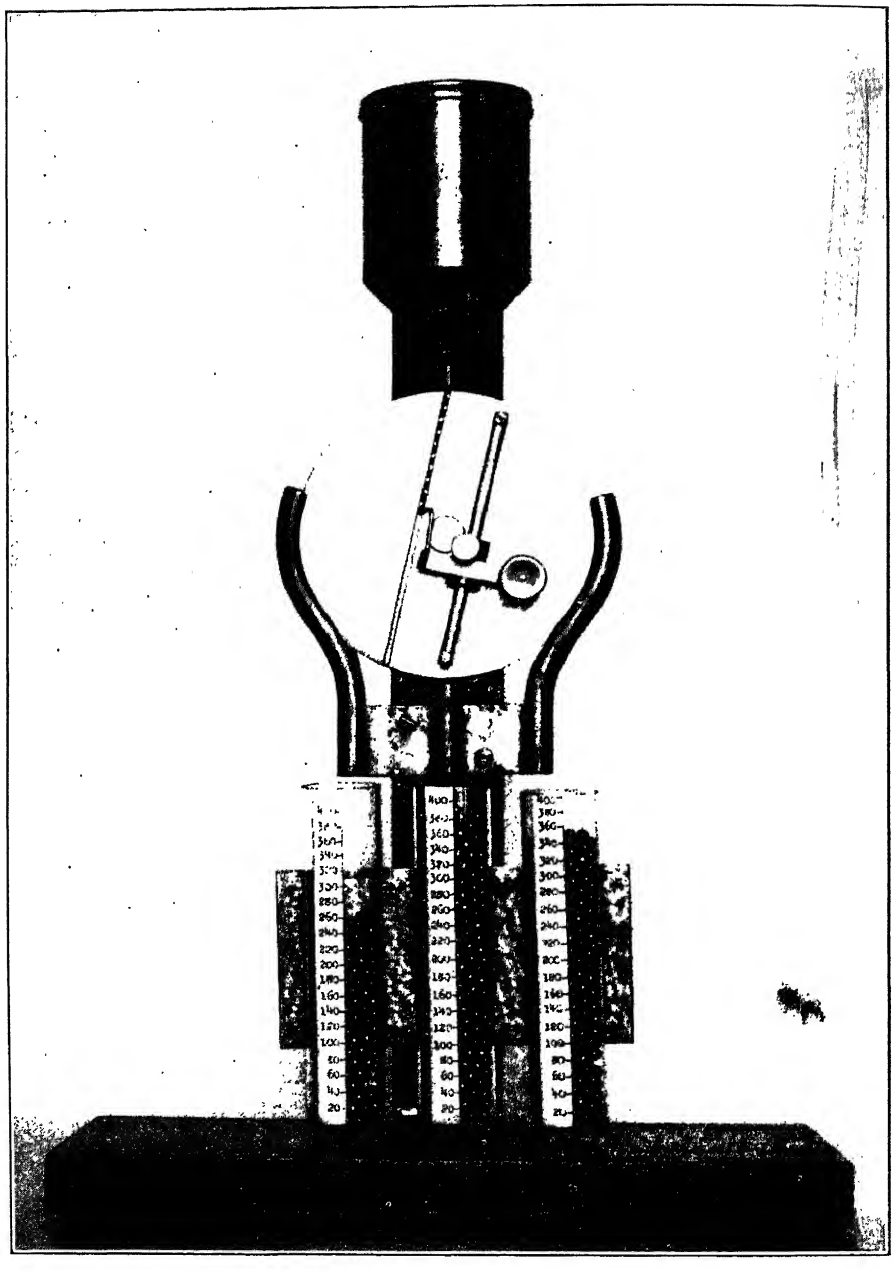


PLATE I—Machine for Sampling of Attributes.

test tube. The photograph shows a typical run. Results are subject to sampling fluctuations, of course, and seldom are 35%, 39%, and 26%; but out of a great many runs, there has never been (and, incidentally could not be) a case in which samples poorer than the lot exceeded samples better than the lot.

A statistical method of estimating lot quality. If c failures out of n have been observed, what conclusion is the practical man to draw, and how is he to avoid a long analysis for which he has no time? These questions are discussed briefly in the following chapter and at length in Appendix A. If the practical man is willing to accept a conservative procedure recommended therein, all he need do is look on Chart 0.5 = I_Q (see pocket on back cover) and read the value of Q_M (estimated lot fraction defective) which is opposite the intersection of the appropriate c curve and the appropriate n line. In the case of $c = 1$, $n = 10$, it is 0.148.

Analysis of some specifications. With the above discussion as a background, the statements about specifications can readily be proved. Calculating the probability that a lot of certain fraction defective will produce any one of a number of kinds of samples at once suggests the process of calculating the probability that a sample of certain fraction defective or less (e.g., that required by a specification) will be produced by any one of a number of kinds of lots. All one has to do is fix n and c, and substitute different values of P and Q in the expression,

$$P_b = \sum_{c=0}^{c} \frac{n!}{(n-c)!\,c!} P^{n-c} Q^c.$$

For a specification requiring that not more than 1 defective shall occur in a sample of 10, this expression becomes $(1 - Q)^n + 10(1 - Q)^{n-1}Q$, which is the probability for not more than 1 defective when the lot fraction defective is Q. The solid line on Fig. 1·2 shows the result of this operation for all possible values of Q; and an examination of this figure, which the author chooses to call the operating characteristic of a specification, will throw a great deal of light on the whole meaning of specifications of this type. Thus, one sees that if lots of articles are submitted which are 10% defective they are not rejected 50% of the time, as might appear from a superficial reading of the specification, but only 26% of the time. A product which is fraction effective 0.95 would generally not be regarded as very good, yet it would pass the specification more than 91% of the time. If it costs more money

to make a product of good quality than of poor quality, it is but natural (barring other influences) for a manufacturer to attempt no level of quality beyond that which makes his cost per accepted article a minimum. Thus, this type of specification renders quality almost exclusively dependent upon margin of profit, which is a very questionable control, because a large margin of profit only encourages poor manufacture, whereas a too small margin of profit may, of course, readily hamper the production of good quality.

Again, noting Fig. 1·2, it is evident that if manufacturer A makes a very good, but not quite perfect product, some of his lots will be rejected; and if manufacturer B presents only very bad lots, some of

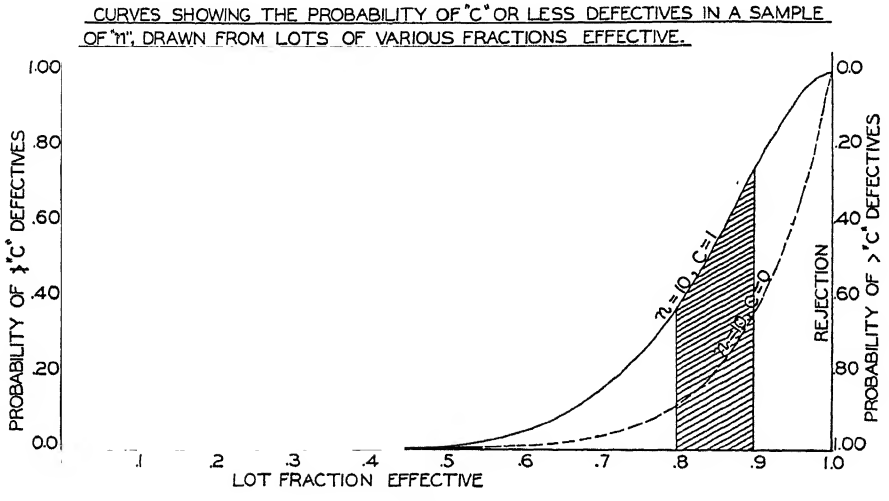


Fig. 1·2

his lots will be accepted. In general, if a large number of lots, all having the fraction defective Q, are presented for acceptance, then $100P_b\%$ of the lots will be accepted and $100(1 - P_b)\%$ of the lots will be rejected. The most discouraging factor, however, inheres not in the rejection of some good lots or in the acceptance of some bad lots but in the fact that *the accepted lots in this case are no better than the rejected lots.* It should be further observed that, if only products of any given level are offered, for example, the shaded area on Fig. 1·2, they will be accepted with occasional rejections. Although no precise statement can be made as to the proportion of accepted product at any given quality level without a knowledge of the number of lots presented at each quality level, nevertheless it is obvious that

the average of the accepted quality will differ little from the average of the presented quality, and that the limits are the same. For the shaded area of Fig. 1·2, and equal frequency of all kinds of lots, the average quality levels are 0.85 (presented) and 0.86 (accepted). It is therefore clearly evident that the quality of product accepted under the old type of specification is practically that which happens to be offered for inspection.

Furthermore, it is obvious that the specification is no reliable key to the quality of products that will be accepted thereunder, and that, if estimate one must, the only rational way to approach the problem is to make some assumption to the effect that a manufacturer who suffers more than some percentage of rejections (say 25%) could not stay in business, and then to take the quality level corresponding to this percentage of rejection as specification quality. Specification quality is to all intents and purposes simply that which happens to be offered for acceptance; and also it appears that the only way to prevent the acceptance of poor quality is by preventing its being offered for acceptance. After this exposition, it appears needless to dwell on the effect of retest provisions which are generally a part of specifications. It is obvious that, since they operate only on rejected lots, they can serve only to reduce still further the level of accepted quality.

Failure of the independent small sample to detect poor quality. One might suspect that these weaknesses could be relieved in some degree by reducing c to 0, thereby allowing no defectives in the sample. This is a stratagem which is frequently employed in practice. The dotted line on Fig. 1·2 is the operating characteristic for the specification requiring 0 failures out of a sample of 10. It is true that it somewhat reduces the chances of acceptance of poor quality; but its penalties fall both on the just and the unjust, and it results in rather high rejections of relatively good quality. Unless a manufacturer attains perfection, a rather fantastic ideal in practice, it is almost impossible for him to have a reasonable assurance of escaping chance rejections. Try what alterations one will with this quite popular type of specification, it is like the poor cooper's barrel: if it does not leak at the spigot, it leaks at the bunghole.

So much for destructive criticism. It is important to know that time-honored methods of making sense out of figures are quite defective in the accomplishment of their objective, uneconomical, and frequently unfair; it is more important to know what can be

done about it. Statistical methods enable one to write specifications which offer reliable assurance of the quality of products accepted thereunder, which are fair to producer and consumer, and which minimize chance rejections of good quality and chance acceptances of poor quality. The method is so logical and simple that it can be outlined in a few brief paragraphs; and, although the discussion will be from the viewpoint of specifications, it will also become evident that by using similar interpretative means one can depend upon small samples in checking the quality of one's product, in sampling consignments of goods, etc. It is believed that the procedure can be most clearly presented by first making just a few observations, then giving an example, and then outlining the underlying principles.

The general situation confronting the sampler. It is rather obvious from the foregoing discussion that an important thing to know is the manufacturer's quality level. This is important for two reasons; first, in order that steps may be taken to correct the faulty manufacturer or stop the influx of his product; and second, because, *without a knowledge of the general quality level of the relatively large stock from which a single lot comes, it is quite impossible to predict even the approximate quality of the lot from the observation of a small sample, even though the sample is perfect.* However, with a knowledge of the general quality level of manufacture, obtained from more extensive sampling, statistical methods (as has already been indicated) frequently enable one to predict the limits within which sampling fluctuations should vary 50% of the time, 90% of the time, or practically all of the time. As long as the results of sampling are within limits of the order of size last enumerated, the variation from sample to sample may be attributed to chance, and no action taken. However, the occurrence of sampling results outside of the limits of chance fluctuations is a strong indication that the manufacturer has changed from his previously established satisfactory level of production.

To illustrate just what is meant without going into the theory behind the procedure, which is briefly explained in Chapter V, a part of a specification governing the velocity dispersion or uniformity of complete rounds of ammunition will be cited. Velocity is measured on a continuous scale; hence this is a case of sampling by variables, whereas the previous discussion was concerned with attributes. However, it is not believed that the change will result in any confusion, and opportunity is afforded for illustrating both types of sampling. The same phenomena previously described still obtain,

viz., the tendency of the small sample to misrepresent the lot favorably (in this case with respect to smallness of dispersion), and the inability of the small sample to furnish an efficient independent basis for inference regarding the quality of the lot. Although average velocity is an important consideration in ammunition, only uniformity or velocity dispersion is being considered at the present time, and the measure of dispersion used is the simplest. It is variously known as maximum dispersion, bracket, or range, and consists merely of the difference between the highest and the lowest value in a series of observations.

The two specifications which must always exist. Prior to writing the specification, a careful study was made of items of this sort, and it was determined that the article could be economically and efficiently manufactured with such uniformity that the *average* range of groups of 5 would not exceed 0.72% of the average velocity. This is taken as specification quality. The aim or goal of the specification, therefore, is to accept practically all lots which are of specification quality or better, and to minimize the chance of acceptance of lots which are poorer. This is the design specification as distinguished from the acceptance specification, and the point wherein the writers of specifications sin most grievously, for it should be obvious that a person is ill equipped to write a specification for a thing unless he has a clearly defined idea of just what is wanted in the first place. The acceptance specification merely defines the quantity and kind of evidence that will be accepted as sufficient that the product will meet the specification goal.[3]

The discussion has proceeded sufficiently far for it to be obvious that the evidence submitted must satisfy three criteria: (*a*) from a relatively large initial sample, it must show that the manufacturer's quality level does not appear to be poorer than specification quality; (*b*) that the character of the product (uniformity) does not appear to be such as to preclude prediction from sample to lot; and (*c*) that by statistical tests there be no indication of a change in quality level as successive small samples are taken from successive lots. In setting up numerical values for these criteria, use will be made of the fact that it can be shown statistically [4] that, if a product is statistically uniform,

[3] These ideas are admirably expressed in "Some Aspects of Quality Control," W. A. Shewhart, *Mechanical Engineering*, December, 1934.

[4] Chapter V explains the basis of this relationship for standard deviation. A similar relationship can be shown to hold for range; see Appendix C.

the range of very few samples of 5 will exceed 2.085 times the average value of this statistic for all samples of 5. Now let us quote the pertinent parts of the specification and see how they operate.

Example of a specification based on statistical methods.

1. *The first lot of a series.* Thirty samples shall be selected as nearly as practicable in the order of manufacture, tested, and the results arranged in 6 subgroups of 5, in the order tested. The range (greatest value minus least) of no subgroup of 5 shall exceed 2.085 times the average range of all 6 subgroups of 5, nor shall the average range of the 6 subgroups times 2.085 exceed 1.5% of the specified velocity.

2. *Subsequent lots of a series.* Five samples shall be selected at random, tested, and the range of the 5 observations shall not exceed 1.5% of the specified velocity.

3. *Assignable causes.* If the range of a sample of 5 representing a lot other than the first exceeds 2.085 times the average range of all previous groups of the same series, steps shall be taken to discover the assignable causes for fluctuation from the standard of the product (these steps are described elsewhere). If, at any time, the acceptance of a lot as prescribed above causes 2.085 times the average range of all accepted lots of a current series to exceed 1.5% of the specified velocity, the next lot submitted shall be considered the first lot of a new series.

Operation of the statistical specification. Figure 1·3 is a graphical exposition of the operation of this type of specification. The first six dots show how the manufacturer's quality level and apparent uniformity of product are established by a relatively large sample from the first lot.[5] Thereafter a running record is automatically kept of the status of the manufacturer's supply to date. In Fig. 1·3 the unit of measure of dispersion has been changed from percentage of velocity to feet per second, as this is simpler and is the usual practice. The individual dots do not purport to show the relative merit of the lots they represent. Statistical methods clearly show that an attempt to distinguish between lots on a basis of such small samples is quite hopeless. They do show, however, that there is no reason for not believing that the lots come from the same level of production;

[5] If sampling were by attributes, much larger samples would be required. The comparison, however, is nonetheless valid.

and the solid central line gives quite a reliable indication of that level to which the whole group belongs. Lot 13 is, of course, a rejection, as such a fluctuation (well outside of the $99\frac{1}{2}\%$ chance limits) cannot be readily attributed to chance. It is important to note that paragraph 3 of the specification calls for action when a point falls outside of the manufacturer's chance limits, even though it is still within the specification limit. This is necessary because such an occurrence indicates lack of statistical uniformity in the manufacturer's product; and, without this control, one is no longer justified in making predictions (i.e., that lots do not appear to be other than essentially the same) based on small samples. The occurrence of a point such as that associated with lot 13 may be due to a change in process, change

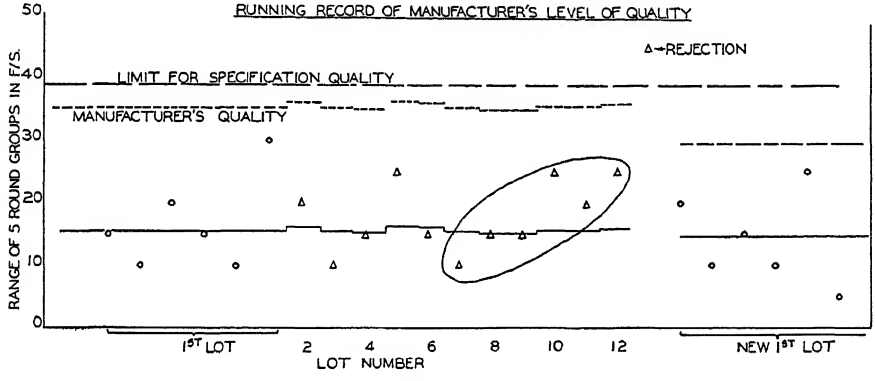


FIG. 1·3.

in raw materials, careless workmanship, sabotage, etc. It is worth while to note that in practice a rejection of this type should not occur, for one of the most valuable features of this type of chart is its ability to predict trouble before it actually occurs. The general trend of lots 7 to 12 inclusive indicate increasing dispersion, and it would be thoroughly evident to one familiar with statistical methods that a rejection would soon occur unless the manufacturer quickly did something to check the tendency for his process to change.

Figure 1·4 shows an approximate operating characteristic for the specification just described. If a manufacturer produces lots of articles which are just barely specification quality, he has a theoretical probability of 0.995 that each lot will pass. This probability is a function of the factor 2.085, which involves some assumptions regarding the variation in the product. Actually the probability of passing

may differ somewhat from this theoretical figure; but, in any event, the manufacturer of just barely specification quality stands a quite reasonable chance of justice. If the product is of somewhat better quality, the probability of passing is even better. Admittedly, the specification allows a possibility of acceptance of lots of poorer than specification quality; but let us examine the nature of this possibility, for it is hedged in two ways: First, the manufacturer has to establish a satisfactory level of production before he has an opportunity of

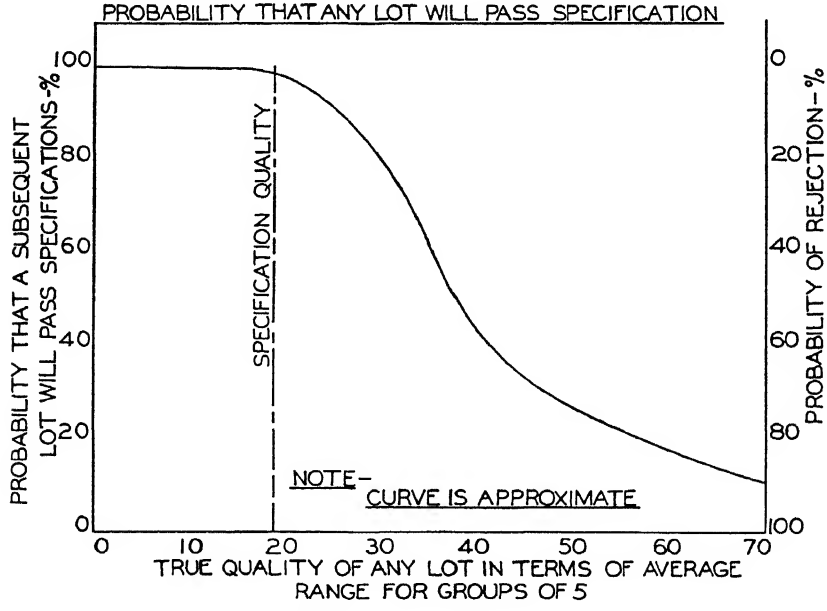


FIG. 1·4.

offering lots for acceptance on the small-sample basis. Second, even though he may by chance secure the passage of one poor lot, the chance of his passing two or three lots becomes increasingly remote; and then, when caught, he must begin all over again with a new first lot of a series. A detailed discussion of the writing of specifications is given in Appendix D.

It therefore appears that statistical methods have yielded several important advantages, even in this simple and rather fragmentary example. A real specification quality has been established which is

certain to be at least fairly closely met by the accepted product. The real quality of lots of the same series (or grand lot), whether of barely specification quality or materially better, will be quite accurately known from the position of the solid central line on the chart. The chance of acceptance of any material proportion of products which are of poorer than specification quality has been made very small; and likewise, the chance of rejection of lots equal to or better than specification quality has been minimized. Surely the good manufacturer would prefer the latter type of specification to the former, for it assures him of a fair deal; but the poor manufacturer would surely prefer the former, which is less rigorous in separating the good from the bad and hence allows him a percentage of successes based on the gambler's chance. A noted financial authority advertises, "If you must play the market, learn the rules." Everyone engaged in any occupation whatever must play the game of probability and chance. It therefore appears expedient to know the rules, and at least protect oneself from the ravages of chance, if one cannot make chance his slave rather than his master.

The case in favor of statistical methods. The avowed purpose of this Introduction was to show the practical man his need for statistical methods, their closely knit relationship with plain common sense, and their ease of application. The need appears to be rather evident, whether one is in the position of buyer or manufacturer, for frequent samples from one's production line, used in the manner outlined, are marvelous aids to economy[6] of process and quality of product. The common sense, however, had perhaps better be presented in summary: (1) The small sample is not sufficient evidence for judging the large lot, hence must be supplemented by other knowledge. (2) This knowledge is readily available in the form of hindsight predicated on past performance. (3) Statistical inference, like many other methods of acquiring knowledge, makes use of the process of postulating a hypothesis and then checking to see whether observed results probably would have arisen under the conditions of the hypothesis. (4) One therefore makes the hypothesis that successive lots are essentially alike; and the small sample, though insufficient as an independent index of quality, is sufficient under these conditions for judging whether or not a change appears to have taken

[6] For a discussion of large savings actually accomplished by means of quality control, see "Statistical Analysis of Metallurgical Problems," by Edward M. Schrock, Supervisor, Statistical Division, Jones & Laughlin Steel Corp., *Metal Progress*, Aug., 1940.

place in quality level. (5) It is then necessary only to determine what the quality level is, either from a single large sample or from the long run of small samples.

As for ease of application, it is not at all necessary for one to compute a whole series of probabilities as was suggested in the development of the first operating characteristic shown in Fig. 1·2. Ordinarily, one is not interested in an exact probability, but only in the fact that the probability is very remote or quite proximate. Hence, for practical purposes a few simple levels of probability such as 0.5, 0.9, 0.995, etc., suffice. Charts of these probabilities have been worked out in advance, so that in application one can read one's data directly from the charts. In short, it may be said that, whereas not all statistical methods are being offered—that would be quite beyond the scope of any moderate-sized book—nevertheless the methods which appear to be of principal importance in engineering and industrial work are being offered, with nothing more in the way of mathematical preparation asked of the reader than grocery-store arithmetic in most instances and high-school algebra in others. Detailed proofs of procedures cited in references or worked out in the appendices or notes are more exacting, of course; but it is presumed that he who wishes to investigate such details will have the moderate additional background to render the task not only light but even entertaining.

CHAPTER II

INSPECTION OF A SINGLE LOT OF ARTICLES—SAMPLING BY ATTRIBUTES

Ce que nous connaissons est peu de chose,
ce que nous ignorons est immense.

—LAPLACE

A method of estimating lot quality. If one merely notes the presence or absence of some quality characteristic in a series of samples and counts how many articles do or do not possess it, the sampling is by attributes. Thus, for example, in sampling a consignment of alloy die castings, one might count those which possess blow-holes and those which do not possess blow-holes. The same consignment might be further classified into those which are corroded and those which are not corroded, those which pass the go, not-go gauge and those which do not, etc. (When measurement is made on a continuous scale such as length, volume, or weight, the sampling is by variables.) Since defects of various kinds are generally of importance, it is convenient merely to speak of those which are defective and, conversely, those which are effective, and allow one to classify defectives as one wishes.

The following discussion refers to small samples from large lots. One of the most frequent problems one meets in sampling is: a certain number of defectives, c, having been observed in a sample of n articles, and nothing else about the lot being known, what inference may one draw regarding the true but unknown proportion of defectives, Q, in the lot; and what probable limits can one assign as the maximum and minimum proportion of defectives to expect in the lot? It was shown in Chapter I that the small sample from a large lot of moderately defective articles tends to misrepresent it favorably, and that the assumption that the lot fraction defective is the same as the sample fraction defective will lead one to an unduly low estimate of the defectiveness of the lot more frequently than to an unduly high one. To correct for this tendency of the small sample to misrepresent the lot favorably, and to yield an answer to this perplexing question,

Chart 0.5 = I_Q has been supplied.[1] This chart makes no assumption regarding homogeneity and is equally applicable to a shipment of electric-light bulbs from Japan, a carefully controlled and highly uniform product from a modern plant, or a heterogeneous mixture of lots: it requires only that the sample be random.[2] Thus, having observed c defective articles in a random sample of n articles and possessing no other information, one finds the intersection of the ordinate erected at the appropriate value of sample size, n, with the curve corresponding to the number of defectives, c, and reads the corresponding value of Q_M. This is a middle-most value of the unknown lot fraction defective, arrived at by considering all possible values of the lot fraction defective as equally likely, finding the probability that each would produce the observed sample, and then selecting the middle value thus found. Hence this is a value which under these conditions will yield an estimate which is too high just as frequently as it is too low. For example, suppose that one has observed 1 defective article in a sample of 50 articles. Looking up the intersection of $n = 50$ with the curve for $c = 1$, one reads $Q_M = 0.0325$, not 0.02, as one would estimate from c divided[3] by n.

[1] Charts of general use are in a pocket at the back of the book.

The unusual nomenclature "Chart 0.5 = I_Q" is adopted because it is peculiarly descriptive of the true nature of the chart. The appropriateness of the symbology inheres in the relationship between the charts and the solution of mathematical functions called the incomplete beta-function ratio on which they are based. The incomplete beta-function ratio has the property of summing the terms of a binomial expansion, and in the symbology here employed would be written $I_Q(c, n - c + 1)$, thereby designating a probability equal to the sum of the last $n - c + 1$ terms of the expanded binomial $[(1 - Q) + Q]^n$ Here are four variables: the probability (which is the value of the incomplete beta-function ratio), c, n, and Q. Values of c, n, and Q, having been fixed upon, the incomplete-beta function ratio yields the value of the probability. However, in practical work one generally prefers to fix the value of the probability and solve for the value of c, n, or Q. That is to say, the inverse solution of the incomplete beta-function ratio is wanted. That is precisely what the charts do. They fix the probability, $I_Q(c, n - c + 1)$, at 0.5, 0.9, 0.995, etc., and offer a solution for any one of the other three variables, c, n, or Q for any fixed values of the other two. Hence, one can solve all possible binomial expansions between $n = 1$ and $n = 500$ for these respective fixed values of the probability, and the charts are rigorous solutions for whatever problems the summations of the binomials are solutions. (Incidentally, only a few of many possible uses are mentioned in this book.) Hence, it is appropriate to designate the charts by the symbols 0.5 = I_Q, 0.9 = I_Q, etc. Plotted values of c and n sometimes differ from their true values by $\pm$ 1 for reasons given in Appendix B.

[2] Random is merely defined as unbiased. Further discussion of the practical meaning of random is given in Appendix A.

[3] The reader is likely to wonder why cognizance is not taken of lot size. Lot size is important: (1) if the withdrawal of a sample with an unusually large or small proportion of defectives can materially alter the proportion of defectives in the remainder of the lot;

The precision of the estimate. However, the estimate of 0.0325 may be higher or lower than the true lot fraction defective, and one undoubtedly wishes some objective means by which he can judge how much the true value is likely to differ from the estimate. This information is supplied by Charts $0.1 = I_Q$ and $0.9 = I_Q$ and also by Charts $0.005 = I_Q$ and $0.995 = I_Q$; but they must be read a little differently, and in accordance with directions printed on the face thereof, because they have been designed to fulfill most conveniently their principal use, which is associated with another problem. Thus, looking up $n = 51$, $c = 1$ on Chart $0.1 = I_Q$, one reads $Q_L = 0.0105$; and the accompanying interpretation is that, under the conditions stated, the probability is only 0.1 that the true Q is less than 0.0105. In like manner, looking up $n = 51$, $c = 2$ on Chart $0.9 = I_Q$, one reads $Q_U = 0.0740$; and the probability is 0.9 that the true Q is less than 0.074.

Thus, having observed 1 defective in a sample of 50, one has three estimates of the true but unknown lot fraction defective: a middlemost value, 0.0325, and two bounding values, 0.0105 and 0.0740, such that the probability is 0.9 that the true value is greater than the former and less than the latter. If these values are too widely spaced, so that the sampler feels that he must have an estimate in which he can have more confidence, then there is just one thing to do: take more samples.

A little experimentation with these charts will soon show that rather larger samples (of the order of 200 to 400) are required to yield estimates with fairly close limiting values. However, lacking evidence other than the sample, there is nothing that can be done about it. There simply is no more information in the sample itself, and any alternative one may choose other than that of utilizing additional evidence is almost certain to prove to be nothing more or less than a comforting delusion. Ways to reduce this huge sample size by logically and mathematically combining existing engineering knowledge with the evidence of the sample will be discussed in Chapter III.

The probabilities 0.1 and 0.9 (or a probability of 9 chances out of 10 that Q is greater than one value and a probability of 9 chances

(2) if the sample tends to exhaust the lot significantly. However, for example, probabilities for a sample of 50 drawn from a lot of 1000, 10,000, or infinity, in general, differ so little that from a practical point of view the distinction can be ignored in drawing small samples from large lots.

out of 10 that it is less than the other) is selected for no reason other than the fact that people are rather accustomed to think in these terms. If one desires higher probabilities, one can use Charts $0.005 = I_Q$ and $0.995 = I_Q$ in precisely the same manner that Charts $0.1 = I_Q$ and $0.9 = I_Q$ were used, thereby obtaining limits (although much wider) for the probability of 199 chances out of 200, or a probability of 0.99 that Q lies between them.

The logic of probable judgments. At this point, it is worth while to reflect for a moment on the nature of probable knowledge. One might ask the question, what are the limits within which Q certainly lies? It does not take mathematics, but only a moment's reflection, to answer this question. No empirical knowledge is ever certain. From the cradle to the grave, one must of necessity act on knowledge which is probable only. If the lot contains N articles, one is not certain of the lot fraction defective Q, even when $N - 1$ have been sampled, for there is still uncertainty regarding the Nth article. Thus, for a sample containing 0 defectives out of 500, Chart $0.5 = I_Q$ gives the estimated value of Q_M as 0.0014, which is small indeed, but which indicates the impropriety of assuming the lot to be perfect. If the sample is from an infinite source, such as successive experiments to determine the acceleration due to gravity, certainty can be associated only with an infinite sample, which, of course, is impossible. The certain limits of Q are 1.0 and 0, and they are certain only because one has defined them that way; those are the limits that one's concept of a fraction prevents one from exceeding.[4] Hence, with the sample size fixed, the width of the probable limits merely varies with the probability; and increased knowledge is available only through increased sample size. The probability level selected, be it 0.9, 0.95, or 0.995, may therefore be unimportant. It is only the interpretation placed upon results which is important, and the 0.9 probability is recommended merely because it appears to be suited to customary habits of thought.

Furthermore, it should be observed that a great deal of nonsense has been written about probability, and the subject has received a great deal of unwarranted abuse, because of failure of people to make a common-sense distinction between cold, mechanical, objective probability and subjective, wish-tinged, personal attitudes. Some writers have called limits such as those just derived confidence limits. They

[4] Reference is made to Chapter X of *Mind and the World Order*, C. I. Lewis, Charles Scribner's Sons, New York.

are nothing of the sort. Confidence is subjective, arises from within oneself, and is subject to a gamut of influences ranging from objective facts to personal idiosyncrasies and unsuppressed superstitions. Thus, for example, two people presented with the same evidence, viz., 1 defective in a random sample of 50, and cognizant of the associated objective probability limits, might feel quite different degrees of confidence, especially, let us say, if one is the buyer and the other the seller. Again, the same evidence and the same objective probability limits might give one quite a different degree of confidence on a morning when it seems that "all's right with the world" from what the same conditions would make one feel on, let us say, the morning after celebrating New Year's Eve. Probability is a most useful tool in aiding the frailty of human judgment, for it is quite impossible for even the shrewd but untutored mind to estimate situations as unerringly and consistently without it as with it; but to fail to distinguish between objective probability and subjective confidence is prejudicial to its intelligent use.

Another error which is perhaps more common and more injurious is the failure of persons to distinguish between probable judgment and indisputable fact. Some go so far as to say that true Q is what it is, and there is no probability about it, with the idea that the statement is at least prejudicial if not a death blow to probability. The statement is absolutely true, but it is not for one moment at variance with probable judgment. The statement that the probability is 0.8 that true Q lies between 0.0105 and 0.074 does not require for its truth that Q actually lies between these limits. It only requires that it be genuinely probable, on the basis of the evidence presented and under the conditions stated.

Probability is always relative to certain evidence and conditions. Having taken 50 additional articles and found 0 defectives, one now has the evidence $n = 100$, $c = 1$; and the corresponding probable judgment yielded by the charts is that the probability is 0.8 that the true Q lies between 0.005 and 0.038. Both statements are eternally true and neither contradicts the other, for each is relative to different evidence. An illuminating example of one's subconscious recognition of the relativity of probable inference to the evidence presented is embodied in the old bridge saying that "one peek is worth two finesses"—an absolutely correct evaluation. *Given all relevant data, a thing is certain. Given anything short of this, the judgment must be probable only, and the probability is relative to the data given.*

The statistical method as a mere aid to judgment. It is believed that the use of the charts as described in this chapter is a valuable aid in arriving at decisions in engineering and industrial work. Neither engineering ability nor business acumen seems proof against the human propensity for overconfidence in predictions based on scant data; and the value of the charts perhaps consists more in showing one what he does not know than in giving assurance regarding the reliability of probable inferences. The rather mild assumption involved in this predictive use, which is briefly described on the face of the charts, is in general on the side of safety. It is a significant condition, however, and, unless the reader is familiar with the subject or does not care for a detailed explanation, he is advised to read the simple but fairly complete exposition of the problem offered in Appendix A.

CHAPTER III

INSPECTION OF A NUMBER OF RELATED LOTS OF ARTICLES—SAMPLING BY ATTRIBUTES

> There was never in the world two opinions alike, no more than two hairs or two grains; the most universal quality is diversity.
>
> MICHEL DE MONTAIGNE: *Of the Resemblance of Children to Their Fathers*

The limitations of sampling evidence. The procedure described in Chapter II had as its objective the extraction of a maximum of the information that exists in the sample itself. Nevertheless, results were meager. For example, suppose that one were presented with two lots of articles, one of which was really 1% defective and the other 5% defective. The charts show that, with sample size 100, the first lot will yield 2 or less defectives 9 times out of 10; and that the second lot will yield 2 or more defectives 9 times out of 10. Thus, it is obvious that even with 100 samples from each of the two lots the probability of distinguishing between them is not so very high; in fact, the probability of rating them in reverse order of merit could scarcely be considered remote.[1] A plot of data from the charts which is made in Appendix A unquestionably shows that for most purposes sample sizes of the order of 300 to 600 are a minimum for achieving satisfactory working estimates of the lot fraction defective from the evidence of a sample, and that very small samples are not only practically worthless for distinguishing between lots (except where defectives are overwhelming), but likely to be positively misleading.

The value of engineering judgment in conjunction with sampling. On the other hand, sampling is expensive. When the sampling process is destructive, as in testing the blowing time of electric fuses, function testing ammunition, etc., large samples are absolutely prohibitive. Practical men know perfectly well that they have long experienced a reasonable measure of success with quite moderate sample sizes, and will accept no such fantastic sample size (as 300 to

[1] The problem of when differences in two samples significantly indicate a difference in the lots sampled is covered in Chapter XI.

600) on theoretical grounds, however sound. What is the basis of this divergence of opinion between the perfectly sound practical man and the perfectly sound theorist? It is not far to seek. The practical man does not confine himself to the evidence presented by the sample. He draws upon his engineering knowledge of the process and conditions relating to the production of the type of article sampled, his knowledge of the character of product made by certain producers, and his experience with similar lots of articles. He knows a great deal about not only what the result of sampling should be, before the sample is taken, but also about whether or not he should pay any serious attention to the test results after the inspection is completed. Undoubtedly he does not use this existing evidence in the most efficient manner, for surely his method is subject to all the frailties and human ills associated with the personal equation; nor has he any scientific and formal process of arriving at his predictions, or of assuring their consistency when made from time to time even under like conditions. Nevertheless, he has a marked advantage over the most precise theorist who is dependent solely upon the evidence presented by the sample.

It has long been felt that an impersonal mathematical approach to the problem, with cogent methods of analysis and scientific procedures, would be productive of even better results than the practical man's moderate measure of success, if only the practical man could put his additional knowledge in a form that the mathematician could use. The practical man's knowledge inheres almost exclusively in an expert judgment that certain lots, classes, consignments, etc., of articles are (for certain reasons known to him) essentially alike. Attempts to combine these two types of special talents were severely hampered because the mathematician wanted something in the nature of an equation, and the type of additional knowledge which the practical man uses never appeared to be conveniently adaptable to equations.

A method of combining practical judgment and sampling evidence. Modern statistical methods, and principally those methods which have been developed since 1924 in connection with the control of quality of manufactured products, offer a simple and effective vehicle for combining the knowledge gained from the sample with existing engineering knowledge, with checks and balances on each so that inconsistency between the sources of knowledge or scantness in either source of knowledge is, in general, promptly revealed, thereby mini-

mizing the probability of erroneous conclusions. The procedure for reducing sample size by means of the application of engineering knowledge is called the grand-lot [2] scheme. It can be outlined in a few paragraphs.

The grand-lot scheme. Suppose that, instead of sampling just one lot, one is sampling a considerable number of lots (either simultaneously or as received). One groups together lots which for engineering reasons are believed to be essentially the same into categories called grand lots. In general, member lots of the grand lot should be of the same manufacturer, made consecutively or at nearly the same time, under essentially the same conditions, etc. Now, let us sample not the lot but the grand lot. The charts introduced in Chapter II place no requirement on homogeneity and will apply [3] whether the grand-lot judgment is good or poor. Furthermore, one can obtain the huge sample size of 300 to 600 required by the conditions of Chapter II by taking only a moderate-sized sample from the respective lots of the grand lot.

Let $n_1, n_2, \cdots, n_n$ be the number of articles in each respective lot sample. Therefore, the sum of the n's, or $\sum_{i=1}^{n} n_i$, is the number of articles in the grand-lot sample.

Let $c_1, c_2, \cdots, c_n$ be the number of defectives in each respective lot sample. Therefore, the sum of the c's, or $\sum_{i=1}^{n} c_i$, is the number of defectives in the grand-lot sample.

By looking up $\left(\sum_{i=1}^{n} n_i, \sum_{i=1}^{n} c_i\right)$ on Chart $0.5 = I_Q$, one has quite a precise estimate of the fraction defective Q_M of the grand lot. If one wishes a measure of the precision, he can look up the corresponding Q_U and Q_L on Charts $0.1 = I_Q$ and $0.9 = I_Q$.

Let it now be tentatively assumed that the respective lot fractions defective are essentially the same as the grand-lot fraction defective.

[2] This term is attributable to Mr. R. H. Kent of the Ballistic Research Laboratory, Aberdeen Proving Ground, who first used it in connection with a study of cannon powder, wherein he facilitated his work by combining similar lots together into a single lot which he called a grand lot.

[3] If the grand-lot judgment is good, i.e., the lots are the same, the sum of the random samples from the lots is the same as random samples from a single grand lot. If the lots are different the sum of the random samples from the lots constitute a non-random sample from the grand lot, but the sample is better than a random sample (see Chapter 19, Yule and Kendall, 11th Ed.). In this event, the charts are a trifle pessimistic.

That was the practical judgment. However, it will be readily admitted that there may be relatively large isolated errors in this judgment. Here Charts $0.9 = I_Q$ and $0.1 = I_Q$ serve their more important function. They instantly and rigorously solve a problem which would otherwise involve a great deal of computational labor and difficulty, viz., the problem of predicting, with a certain weight or probability, the kind of sample a given lot will produce. Merely by entering the charts with the grand lot Q and the lot sample size n, they yield two values of c, $c_{U0.9}$ and $c_{L0.1}$, such that, if the lot fraction defective is really Q, the probability is 0.9 that a random sample of n would not have contained more than $c_{U0.9}$ defectives and the probability is 0.9 that it would not have contained less than $c_{L0.1}$ defectives. A lot whose c is greater than $c_{U0.9}$ is suspected of being poorer than fraction defective Q, for otherwise the probability is 0.9 or greater against this occurrence. A lot where c is less than $c_{L0.1}$ is suspected of being better than fraction defective Q, for otherwise the probability is 0.9 or greater against this occurrence. Lots whose samples fail to meet this criterion are termed suspected mavericks, for they are suspected of not being of essentially the same quality level as the grand lot.[4]

Test conditions are generally such that only one class of these suspected mavericks should be subjected to a more extensive test in order to determine their quality more precisely, for, if the grand lot is satisfactory, lots which are better than the grand lot are satisfactory. If the grand lot is unsatisfactory, lots which are poorer than the grand lot are unsatisfactory. The lots whose samples fall between $c_{U0.9}$ and $c_{L0.1}$ are accepted as bona fide members of the grand lot and tentatively assigned the grand-lot fraction defective. Thus the mathematical method works hand in hand with and checks and supports the practical judgment.

Test of the grand-lot judgment. However, there is still one more logical check to be made. There are experts and experts, and some judgments may be rather poor. One must be saved from rash judgment. By entering Chart $0.1 = I_Q$ with Q equals 0.1 (the probability basis of the chart) and n equals the number of lots in the grand lot, one reads a value of c such that, if the respective lots are really of fraction defective Q, the probability is 0.9 that not more than this number, c, of suspected mavericks would have

[4] It should be noted that the probability is not 0.9 or any other assignable figure that a suspected maverick is greater or less than the grand-lot level. This would be a most misleading inference from so-called inverse probability.

occurred above $c_{U0.9}$. (The same criterion applies to the number below $c_{L0.1}$, but it is recommended that the test be applied to only one limit.) It is believed that the logical basis of this test is obvious, for it amounts merely to assuming that the grand-lot hypothesis is true, and then checking to see whether observed results would have probably arisen under the conditions of the hypothesis.[5] If the grand-lot judgment meets this criterion, the lots other than the suspected mavericks are assigned fraction defective Q_M; otherwise the grand-lot judge is called upon to revise his grand-lot grouping, if a logical basis for regrouping exists, or the grand lot must be abandoned and resort made to individual sampling.

To one of a reflective turn of mind the question will arise whether the grand-lot fraction defective Q_M should be altered in view of the elimination of the suspected mavericks. Obviously, one would like to include in the estimate of Q_M all suspected mavericks which are not really mavericks, and exclude all others. A working rule to accomplish approximately this end is offered. Enter Charts 0.005 $= I_Q$ and 0.995 $= I_Q$ with the grand lot Q and the lot sample size n, thereby obtaining two more values of c, $c_{L0.005}$ and $c_{U0.995}$, such that, if the lot fraction defective is really Q, the probability is 0.995 that the sample would not have contained less than $c_{L0.005}$ defectives and the probability is 0.995 that it would not have contained more than $c_{U0.995}$ defectives. If any suspected maverick contains more than $c_{U0.995}$ or less than $c_{L0.005}$ defectives, it should be eliminated from the grand lot, and the grand-lot fraction defective Q_M recomputed. However, it is not generally necessary to repeat the remainder of the analysis. The above action is necessary to keep an isolated error of extreme character from unduly prejudicing the whole analysis.

Illustration of the grand-lot scheme. By way of illustration of the operation of the grand-lot system, consider the following data from the inspection of a tentative grand lot:

Lot	n	c	Lot	n	c
1	40	1	7	40	3
2	40	2	8	40	1
3	40	2	9	40	7
4	40	5	10	40	2
5	40	0	11	40	1
6	40	4	12	40	1
				$\Sigma n = 480$	$\Sigma c = 27$

[5] The probability is specifically confined to an "if then" proposition: if the lot is of the grand-lot level, then the probability is 0.9 against this occurrence. Whether the observed sample represents a remaining 0.1 chance or some other chance is quite unknowable.

Looking up $n = 480$, $c = 27$ on Chart 0.5 = I_Q, one reads the estimated lot fraction defective $Q_M = 0.057$. Looking up $Q = 0.057$, $n = 40$ on Chart 0.9 = I_Q, one reads $c_{L0.1} = 1$. The same coordinates on Chart 0.1 = I_Q give $c_{U0.9} = 4$. Thus, if a lot is really fraction defective 0.057, the probability is 0.9 that it will not yield less than 1 defective in a sample of 40 and the probability is 0.9 that it will not yield more than 4 defectives in a sample of 40.

A plot of the data is shown in Fig. 3·1. Lot 5 is suspected of being less defective than the grand lot. Lots 4 and 9 are suspected

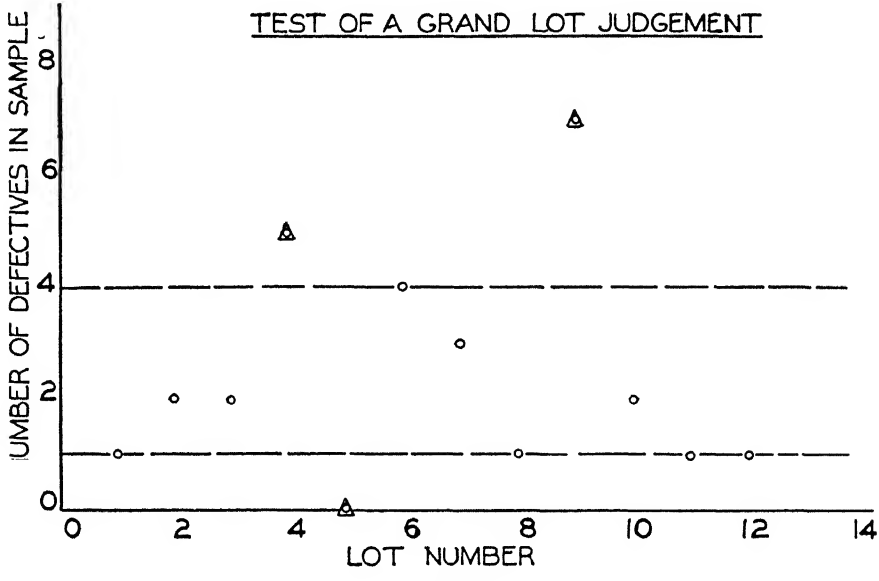


FIG. 3·1.

of being poorer than the grand lot. All the other lots are tentatively assigned fraction defective 0.057.

Looking up $Q = 0.1$, $n = 12$ on Chart 0.1 = I_Q, one reads $c = 2$. Thus the probability is 0.9 that, if the 12 lots are essentially the same, not more than 2 suspected mavericks would occur above $c_{U0.9}$. The grand lot meets this criterion, and lots other than 4, 5, and 9 are assigned fraction defective 0.057.

If this is a satisfactory grand lot, it is not necessary to retest lot 5 for it appears to be even better. Lots 4 and 9 should be retested with a larger sample size (generally the grand-lot sample size), to confirm their quality.

An appraisal of the grand-lot system. This system appears to estimate quickly, simply, economically, and reliably the lot fraction defective and to save as much as about 80% of the theoretical sample size required for single lots. Very good grand-lot judgments are desirable but not essential to the operation of the system, as very poor ones will almost invariably be caught. Poor grand-lot judgments result in retesting more suspected mavericks and serve to decrease the efficiency of the system. Of course, if one ignored engineering judgment, and deliberately juggled lots so as to meet criteria (thereby nullifying the practical man's contribution), the system would be reduced to mere sampling of single lots on a small-sample basis.

The minimum grand-lot sample size and the minimum lot sample size are contingent upon the precision desired. In sampling by attributes the question of sample size is amazingly free of complex considerations. The lot sample size must be such that lots of significantly poor quality have a good probability of being detected. Manifestly, an inferiority in quality cannot be detected unless the sample exhibits at least 1 failure. Hence, having decided on engineering grounds what constitutes a significant deviation in quality—e.g., that lots which are of fraction defective 0.04 or poorer should be classified separately from those between 0.04 and perfection—one needs only select a lot sample size, n, such that lots of fraction defective 0.04 have a probability, P, of yielding at least 1 defective in a sample of n. For the probability 0.9, and $Q = 0.04$, the corresponding sample size, n, is readily read from Chart $0.9 = I_Q$ as $n = 54$. For other probabilities one need only read the other charts in the a priori sense. For convenience, Fig. 3·2 gives this type of information for the probability 0.9 for a considerable range. In Chapter X, under tests of increased severity, methods will be shown for reducing the sample size even further.

In general, the grand-lot sample size is dictated by some action limit. For example, if one were going to scrap, rework, or 100% inspect lots of articles which appear to be poorer than $Q = 0.08$, one would wish a reasonable probability that grand lots of articles which were significantly above this level, say $Q = 0.06$, would not by chance be rated as poorer than this action limit. Figure 3·3 shows a chart for minimum grand-lot sample size predicated on a probability of 0.9 of not misgrading lots which are three-fourths as good as the action limit. This type of chart can be constructed for a variety of conditions by entering charts I_Q with initial data and plotting results. It

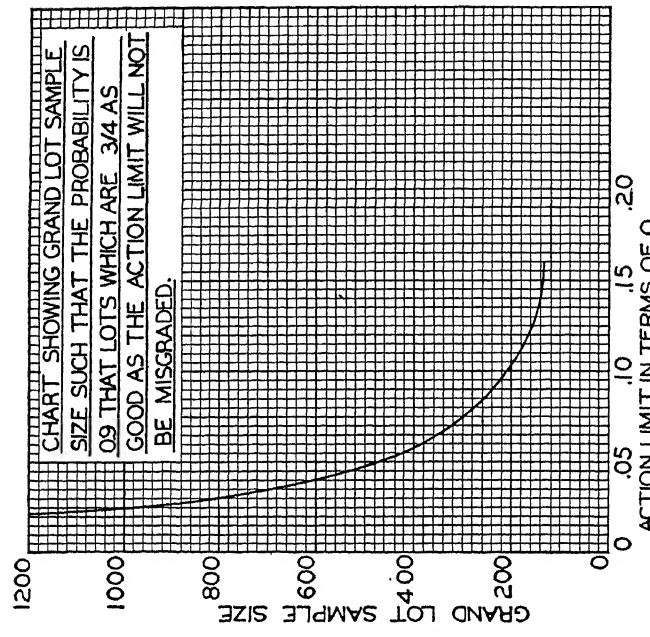


CHART SHOWING GRAND LOT SAMPLE SIZE SUCH THAT THE PROBABILITY IS 09 THAT LOTS WHICH ARE 3/4 AS GOOD AS THE ACTION LIMIT WILL NOT BE MISGRADED.
GRAND LOT SAMPLE SIZE
1200
1000
800
600
400
200
0
0
.05
.10
.15
.20
ACTION LIMIT IN TERMS OF Q

FIG. 3·3.

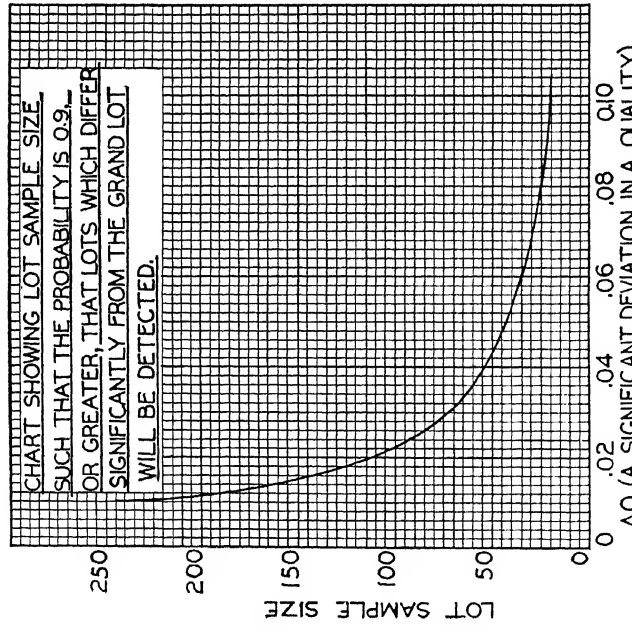


CHART SHOWING LOT SAMPLE SIZE SUCH THAT THE PROBABILITY IS 0.9 OR GREATER, THAT LOTS WHICH DIFFER SIGNIFICANTLY FROM THE GRAND LOT WILL BE DETECTED.
LOT SAMPLE SIZE
250
200
150
100
50
0
0
.02
.04
.06
.08
0.10
ΔQ (A SIGNIFICANT DEVIATION IN A QUALITY)

FIG. 3·2.

may be remarked that, a considerable part of the time, the lot sample size will force a grand-lot sample size of equal to or greater than the minimum.

One may well question what one should do when there is not a considerable number of lots on hand but when instead one is receiving consecutive lots from a manufacturer. The answer has already been implied by the discussion in the introduction. From the first few lots, or at least from the first lot, one must take a large sample in order to have a reliable estimate of the manufacturer's general level of quality as measured by the fraction defective Q. From then on, one can treat his successive lots as additional members of a grand lot, testing each suspected maverick by a large sample to see if its quality is really satisfactory. However, the occurrence of an extreme suspected maverick (beyond the 0.995 or 0.005 limits) or an excessive number of suspected mavericks should result in terminating the manufacturer's grand lot and in making him qualify all over again.

It appears desirable to observe in closing this chapter that the system outlined offers little that is novel. The procedure followed is one of the most time-honored and reliable methods for gaining knowledge, viz., the *postulating of hypotheses*, making assumptions, *and then testing to see whether or not observed results would probably have arisen under the conditions of the hypotheses.* The hypothesis that certain groups are essentially alike, is tantamount to a concise way of summarizing the practical man's existing knowledge. The operational procedure is frankly an adaptation of the principles of quality control, which has been thoroughly tried in industry and is admirably described in W. A. Shewhart's, *Control of Quality of the Manufactured Product.*

CHAPTER IV

PROCESS INSPECTION—SAMPLING BY ATTRIBUTES

> Dead flies cause the ointment of the apothecary
> to send forth a stinking savour· · ·.
> —*Ecclesiastes*, Chapter 10, Verse 1

Advantages of statistical quality control. The inspection processes described thus far presumably were conducted after manufacture. Their objective was either the rating of lots of articles in order of merit, or the acceptance or rejection of lots. Process inspection has to do with the control of quality during manufacture. Incidentally it yields a highly reliable index of quality, but it has as its primary objectives (1) uniformity of product, (2) catching trouble—i.e., detecting a change in product due to some trouble in process before it has progressed to a serious stage—and (3) saving the manufacturer money by helping him to do what he wants to do more efficiently and economically.

For one who has any experience in the supervision of manufacture it is unnecessary to dwell upon the vexing problem which arises when everything just seems to "go sour" with the process, when the product is terrible, rejections are occurring, and one damns everything from the foremen to the raw materials, but does not know where to put his finger on the trouble. Under a system of quality control this situation could hardly arise; but if it did, one would have some excellent leads regarding the causes of the trouble. Another plague which drives executives to early graves is the discovery of an unusual but serious error in manufacture, and the concomitant tragic wonder which arises regarding how many hundreds or thousands of the product may have gone out to service before the error was discovered. Quality control not only detects a change in quality promptly, but frequently predicts the approach of trouble before it actually occurs. Process inspection under a system of quality control gives better inspection at less cost, effects a marked economy by reducing rejections, gives one an indubitable guarantee of the quality of the product,

indicates ways of improving quality at no increased cost, and brings to light much engineering knowledge regarding the process. But the advantages of statistical methods are at most merely of incidental interest in a book on methods and their application.

It is not proposed to discuss in detail the theory of quality control, the details of its manifold advantages, or even, thoroughly, its mode of application. For these matters the reader is referred to Shewhart's original work [1] on the subject or to a simplified and condensed version [2] of quality control prepared under the joint sponsorship of the American Society for Testing Materials and the American Society of Mechanical Engineers. It is proposed to give only the irreducible minimum of simple directions essential to the practical application of these powerful methods to some of the commoner problems, in the solution of which it has proved of greatest economic value, and to offer a few simplifying and labor-saving shortcuts for facilitating its use by persons more skilled in their particular fields than in mathematics or statistics.

Conditions favorable to sampling by attributes. It has previously been remarked that sampling of variables is in general more efficient than sampling of attributes; i.e., one gains more information from the inspection of a given number of articles. However, it sometimes happens that it is cheaper to sample a large number of piece parts by attributes (say by a go, not-go gauge or by passing articles over two weighing scales set at certain tolerance limits) than it is to sample a much smaller number by quantitatively measuring the variable linear dimension, actual weight, etc. Again, it frequently happens that no quantitative way of measuring a thing is known. For example, the small primer which is assembled into the base of complete rounds of ammunition appears merely to either fire or not fire when struck by a given blow of a firing pin. Sometimes there is no convenient unit of measure of a quality characteristic; [3] for example, in producing plated or painted surfaces, photographic film, etc., one may resort to merely counting surface defects per unit area. Hence process inspection by sampling of attributes cannot be omitted, even though sampling of variables may offer more attractive results,[4] is more frequently met in practice, and is in general preferable.

[1] *Op. cit.*

[2] 1933 *A.S.T.M. Manual on Presentation of Data*, Second Printing, March, 1933, A.S.T.M., Philadelphia.

[3] This condition is given special treatment in Chapter VIII.

[4] See Chapters VII and VIII.

Similarity of process inspection to the grand-lot scheme. When process inspection is on a basis of sampling by attributes, the statistical method known as quality control operates almost precisely like the grand-lot scheme described in Chapter III. It is worth recalling that the inspection system was drawn as a parallel to quality control. The whole flux of continuous product corresponds to the grand lot; the successive samples from increments of the continuous output (call them sublots, batches, etc.) correspond to the samples from the member lots of the grand lot. The hypothesis is to the effect that the quality of the product remains essentially the same from increment to increment; and samples which fall beyond the probable limits correspond to the suspected mavericks and are indications of trouble in the process.

The basic theory of quality control. The logic underlying quality control is almost basic in its simplicity. Since the beginning of time no two things appear to have been made exactly alike. Products vary from article to article. Variability in the manufactured product can be classified under two heads: (1) that which is due to the operation of a number of systems each composed of a large number of chance causes, and (2) that which is due to assignable causes.[5] The variability due to chance causes is inherent in the system of manufacture and cannot be materially altered without significantly changing the system of manufacture. The variability due to assignable causes can be eliminated by discovering the assignable causes and removing them. As long as the variability of the product is due only to the operation of chance causes, the product will follow the laws of chance associated with the causes. Consequently, knowing these laws, at least approximately, one can predict with corresponding proximity the limits within which any assigned percentage of samples of size n should lie. However, when an assignable cause of variability enters the system, it will superimpose an additional variation upon the chance variation, thereby causing samples to exceed the probable limits. This is a parallel to the suspected maverick. Hence, the statistical method will show the presence of an assignable cause of variability.

Since samples are generally taken at regular intervals either with

[5] Shewhart designates these as Assignable Causes Type I, and subdivides the chance causes into two categories, one of which he designates as Assignable Causes Type II, which consists of a system of chance causes, having a predominant effect, which can be discovered, and at times eliminated with more or less concomitant change in system. The author does not wish to go into this refinement in this book.

respect to time or to multiples of units of the product, the method frequently gives a good indication of the nature of the assignable cause. It cannot be too strongly stressed that *order of occurrence is of great importance in this type of sampling;* i.e., the sampling results must be charted in the order of their occurrence. Once the order is lost and the results scrambled, all the king's horses and all the king's men cannot put the data back together again. The heart and soul of the system consists in detecting whether samples taken not randomly, but in some purposive order such as the order of production, meet criteria associated with random samples. If so, it follows that arbitrary order had no effect on sampling results, and hence that the product did not appear to change with respect to a parameter such as time; i.e., there is no reason for not believing that the product is statistically uniform. Is that not simple and logical?

Order of occurrence is also helpful in indicating trends and giving important leads regarding the causes of unwarranted variation. The actual location and elimination of the assignable causes of variation constitute an engineering problem. The above information, however, is all the engineer either needs or desires. The principal source of power in the system may be regarded as its intelligent use of hindsight, since it considers not only the sample at hand but also its predecessors. It should be remembered that, although the small sample does not offer sufficient evidence for judging quality efficiently, it does offer a sound basis for judging whether or not the increment sampled appears to be essentially the same as its predecessors.

The relationship between cogency of hypothesis and probability level. Again noting the parallel to the grand-lot scheme which was drawn from quality control, one observes that his hypothesis in the present case—viz., that the successive increments of processed product are essentially the same—is much stronger. Consequently, it takes a higher probability to cause one to abandon one's hypothesis, i.e., suspect real variation from the established level of quality. The upper and lower limits of allowable variability in the number of defectives, c, in a sample of size n which are generally taken in this type of work are the mean number of defectives per sample, $\bar{c}$, plus or minus three standard deviations of c, σ_c, where a standard deviation of c is computed as, $\sigma_c = \sqrt{Q(1 - Q)n}$. The probability associated with limits computed in this way is very high for the upper (of the order of 0.995) and very low for the lower (of the order of 0.005), but varies considerably with different values of Q, n, and c. This

method has given splendid results in practice, and it is therefore more with a view to eliminating the labor of multiplying out Q times $(1 - Q)$ times n and taking the square root of the product than to remove a slight variation in probability that Chart 0.995 = I_Q and Chart 0.005 = I_Q are supplied from which the limiting values of c can be read directly. It is believed that an illustrative example will clear up any doubts about the mode of operation.

An illustrative example. Figure 4·1 shows the results of test of samples of size 40 taken from the production process of a component

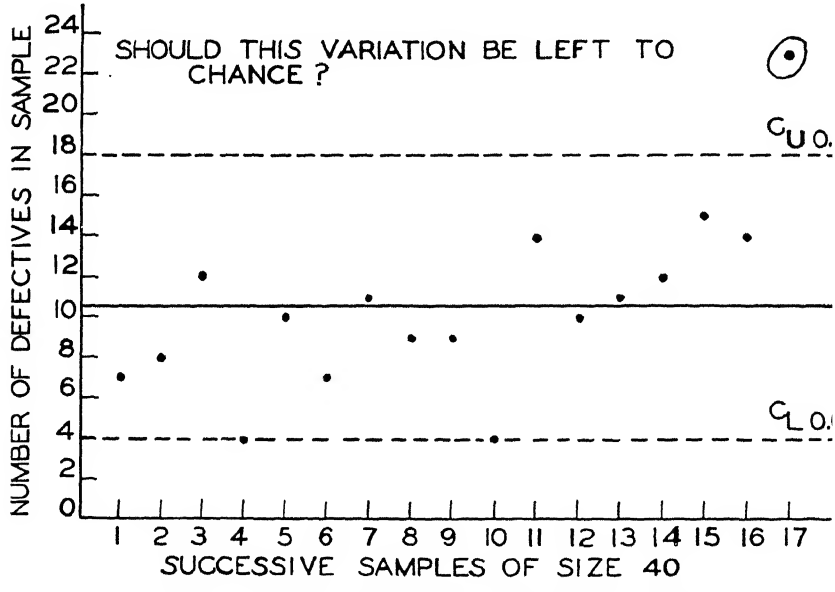


FIG. 4·1. Quality Control Chart—Sampling by Attributes.

mechanism of a powder train anti-aircraft time fuze.[6] The plot shows the number out of each sample of 40 which failed to function (defectives) under the conditions of the test.[7] The dotted limits shown in Fig. 4·1 are predicated on the first 17 samples of 40, consisting of a total of 680 articles and showing a total of 180 defectives.

[6] As a statistical illustration, the chart is like the original, but the data have been altered so as to render it entirely fictitious.

[7] The large sample fraction defective does not necessarily indicate poor quality, for the test conditions were made more severe than the working conditions; otherwise only slightly defective batches would be likely to show no failures at all. The principles of this type of testing are given in Chapter X.

Chart 0.5 = I_Q goes only to $n = 500$. For samples of larger than 500 one may just as well use the average sample fraction defective for Q; i.e., $\frac{180}{680}$ or 0.2647. Entering Chart 0.005 = I_Q with $Q = 0.2647$, $n = 40$, one reads [8] the nearest c line as 18. The probability is 200 to 1 against observing more than 18 defectives in a sample of 40 from a lot of fraction defective which is equal to the average of the 17 samples. Hence, if sample 17 came from the average quality level, the probability is extremely remote that it would have had 23 defectives. It is highly probable, therefore, that an assignable cause of variation is operating to produce variation beyond that which should be left to chance. One should note, however, that one cannot say how probable this contingency is (see footnote regarding suspected mavericks, page 29). Actually, there was an assignable cause introduced between samples 10 and 11, and the lack of control shown by sample 17 might have been at least suspected from the upward trend (which could happen by chance) of samples 11 to 16.

If the information yielded by this chart appears meager, it is at least highly probable knowledge and provides a sound basis for action, which is appreciably more than one could say for unaided judgment based on the mere observation of an apparently somewhat high sample. However, it will be shown in Chapters VI and VII that a similar system based on sampling by variables appears to offer considerably more guidance. It was remarked in the beginning that sampling by attributes, although simple and cheap, has its disadvantages.

Initial conditions for quality control. In contemplating the initiation of a system of process inspection based on quality control, one naturally wonders how many points one should have initially before looking up limits, and how frequently limits should be reassigned. The answer to such a question involves engineering and economic considerations associated with the particular type of production sampled, and hence does not readily admit of a general answer. All statistical methods are merely tools in the hands of the engineer and must be used with the same good judgment and common sense as other principles, methods, and tools. At least a partial answer to this question, however, will be found in the example (taken

[8] An approximation is involved here. The probability 0.005 does not correspond to a whole number of failures. In sampling of attributes, however, one cannot observe a fractional failure; hence the only practical thing to do is to take the nearest whole number. This results in varying the probability slightly from 0.005.

from actual working conditions) of a simple quality-control system given in Appendix C. In this example, all necessary details of the quality-control system, including the number of points on which limits should be based, are presented in only five pages, and so simply that foremen and inspectors having no more than high-school education successfully put it into operation without additional instruction.

CHAPTER V

INSPECTION OF A SINGLE LOT OF ARTICLES—SAMPLING BY VARIABLES

Great fleas have little fleas upon their backs
to bite 'em,
And little fleas have lesser fleas, and
so ad infinitum.
—De Morgan: *A Budget of Paradoxes*

The concept of frequency distribution. A discussion of sampling of variables necessitates at least a rudimentary conception of frequency distributions. However, such a conception is neither difficult nor a matter of mere academic interest. Anyone either directly or indirectly associated with repetitive process (either mass production or repeated experimental observations) will find the concept a most helpful background in visualizing and more completely understanding his daily work.

To begin with, let it be observed that try as one will one cannot produce any two things which are exactly alike. Suppose, for example, that one were given a considerable length of dowel rod and told to cut one hundred 2-inch lengths. Some would perhaps be 1.98 inches, some 1.99, some 2.01, and perhaps some, let us say, as far off as 1.95, variation depending upon skill, kind of tools, etc. However, the majority of the pieces would be closely clustered about some average length, say 1.97 inches; and the number which occurred of any given length would be less, as the given length diverged more from the average. If a large number were cut, the distribution of lengths would presumably be somewhat of the form shown in Fig. 5·1, where the width of the bars represents the limits within which lengths of rods are grouped, and the height of the bars shows the number in the group. That is to say, small deviations from the average occur frequently, and large ones relatively infrequently. The diagram is illustrative of a frequency distribution, and it is often quite helpful to think of a lot of articles in terms of a frequency distribution, rather than as merely being between certain limits such as would be determined by a go, not-go gauge.

A smooth curve through the centers of the tops of the bars of Fig. 5·1 is representative of what is variously known as the error law, the Gaussian law, the probability curve, or normal distribution.[1] It is shown in Fig. 5·2, where the average is represented as $\bar{X}$ (the average of the measured characteristic, X), and the dispersion of the individuals about the average is measured in standard deviations. The standard deviation is conventionally represented by the Greek letter, σ. The equation of this curve is well known, and under normal law, plus and minus 3σ includes practically the whole universe (theoretically 99.73%). Plus and minus 0.6745 includes 50% of the total X's—hence the familiar term probable error (the 50% zone), i.e., the error which is as likely to be exceeded as not.

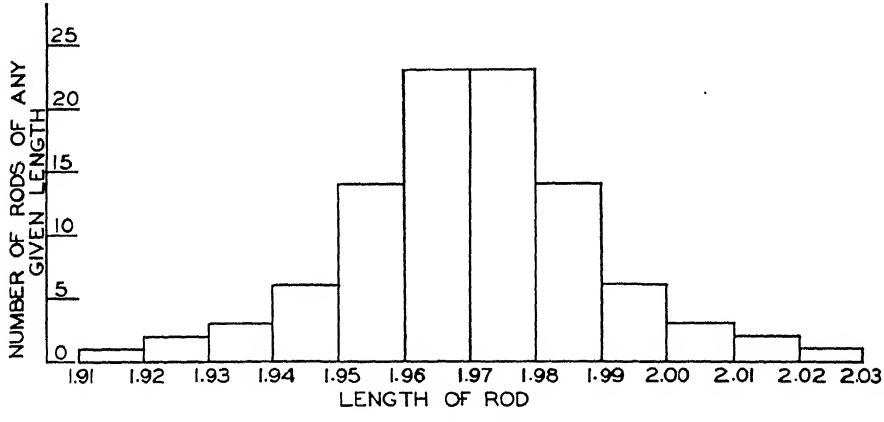


FIG. 5·1. Schematic Frequency Distribution of Lengths of Rods

The average and standard deviation. It is thus apparent that when one samples any lot, universe, or population he is, in general, either consciously or unconsciously seeking estimates (a) of the central tendency (the average, $\bar{X}$), and (b) of a measure of the dispersion of the individuals about that center (the standard deviation, σ), whether he thinks in these terms or chooses others. Knowing the average and the standard deviation, one would know everything about the whole distribution, if only it were normal.

The assumption of normal law is a most popular and comforting simplicity. In practice, distributions are at best somewhat skewed,

[1] There are many other well-known distributions. For example, the work of Chapters II, III, and IV involves the Bernoulli distribution; that of Chapter VIII is predicated upon the very skewed Poisson distribution.

flattened, or peaked, thereby playing havoc with precise percentage zones, derived in the conventional manner. Yet, almost every time one reads an empirical constant together with a precision measure such as 324.074 ± 0.006, where the 0.006 is quoted as the probable error, the assumptions have been made (*a*) that the universe is normal and (*b*) that the estimate of the dispersion (P.E. or std. dev.) is correct. These assumptions are practically never justified and lead to grave error, as will become apparent as the practical procedure about to be offered is developed.

Certainly one of the first requirements of a practical procedure is that it must not purport to do what it cannot do. Furthermore, the assumption of normality is generally not necessary. First, let it be frankly admitted that in practical work one can never have precise

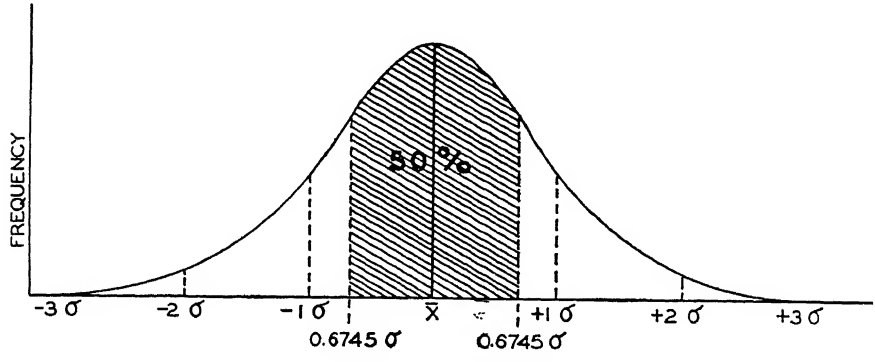


FIG. 5·2. Schematic Diagram of a Normal Universe.

probable limits. Is that a serious fault? In general it is not, for in practical work it generally does not matter whether the probability is 0.9973, or 0.95, just so it is sufficiently proximate or sufficiently remote. Even if the admission of approximateness were a fault, it would scarcely be the part of wisdom to exchange it for a spurious exactitude. A simple chart method will be given for quickly arriving at reliable probable limits; but instead of attaching any exact probability to the limits, they are frankly designated as fairly probable (of the order of 0.75 to 0.90), highly probable (of the order of 0.89 to 0.95), and extremely probable (of the order of 0.95 to 1.0). Only one mathematical procedure will be asked of the reader, and that procedure consists of the relatively simple process of calculating the standard deviation. Even this calculation could be avoided in many cases by use of a method outlined in Chapter XII and Appendix C,

but it is of such great fundamental importance that it should by no means be omitted.

Computation of limits for $\bar{X}$ and σ. Given a set of observations $X_1, X_2, X_3, \cdots, X_n$, the sample standard deviation, σ, is calculated by taking the average of these observations; subtracting the average from each of the observations, thereby obtaining a series of deviations from the average; squaring these deviations; taking the mean of the squared deviations; and, finally, taking the square root of this mean. For example, suppose that, during the cutting of a large number of nominally 2-inch lengths of dowel rod, samples were taken from time to time, until 20 samples were accumulated. Upon careful measurement, the 20 samples proved to be as tabulated in the column X. The remainder of the tabulation shows the calculation of the average and standard deviation.

COMPUTATION OF STANDARD DEVIATION

X	$(X - \bar{X})$	$(X - \bar{X})^2$	X	$(X - \bar{X})$	$(X - \bar{X})^2$
1.95	0.02	0.0004	1.98	0.01	0.0001
1.99	0.02	0.0004	1.95	0.02	0.0004
2.01	0.04	0.0016	1.95	0.02	0.0004
1.96	0.01	0.0001	1.96	0.01	0.0001
1.95	0.02	0.0004	1.96	0.01	0.0001
1.97	0.00	0.0000	1.95	0.02	0.0004
1.96	0.01	0.0001	1.97	0.00	0.0000
1.98	0.01	0.0001	2.01	0.04	0.0016
1.95	0.02	0.0004	1.99	0.02	0.0004
1.98	0.01	0.0001	1.98	0.01	0.0001

$$\sum_{i=1}^{n} X_i = 39.40 \qquad \bar{X} = 1.97$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = 0.0072$$

$$\sigma = \sqrt{\frac{0.0072}{20}} = 0.019$$

That is to say,

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}}$$

Now, let it be observed that, although small samples do not yield an average which tends to differ consistently from the lot average (as small samples by attributes tend in general to yield consistently a smaller fraction defective than the lot fraction defective), nevertheless the average based on a small sample may obviously be greater than or less than the true lot average, owing to the influence of chance in the selection of samples. One therefore wishes some estimate of the limits within which the true lot average probably lies. These limits are given by the Alignment Chart for Reading the Probable Range of Error in an Observed Mean. In the present instance, the observed mean, $\bar{X}$, is 1.97; the observed standard deviation, σ, is 0.019; and the sample size, n, is 20. By placing a straight edge on the chart and lining up 20 (the sample size) on scale A with 19 (the significant digits of the observed standard deviation) on scale B, one reads 91 and 136 on scales C and D, respectively. Placing the decimal point by inspection,[2] we may therefore say that it is highly probable (probability of the order of 0.90 to 0.95) that the lot mean lies between 1.97 ± 0.009, and that it is extremely probable [3] that the true lot mean lies between 1.97 ± 0.014. This statement graphically transmits a great deal of information about the average of the lot. If one wished the probable error of the mean (or theoretically 0.50 probable limits), it could be readily computed by multiplying 0.0136 by 0.2248, thereby obtaining the answer 1.97 ± 0.003. However, probable error is not theoretically correct, except in the classical normal distribution; and, from the practical viewpoint, it seldom offers sufficient assurance to be a satisfactory guide to action. Hence, the limits offered by the chart are believed to be preferable.

Interest is as frequently centered in the measure of dispersion as in the measure of central tendency. Therefore consider the observed value of the standard deviation, 0.019. It can be shown that the standard deviation calculated from a small sample tends to be smaller than the true standard deviation of the lot. (This is analogous to the general tendency of the small sample by attributes to misrepresent

[2] If n is between 9 and 100, the reading on scale D is less than the observed standard deviation, and lies between σ and 0.3σ. If n is less than 10, the reading on scale D is between σ and 1.7σ. No other rules for pointing off are necessary. The reading on scale C is always $\frac{2}{3}$ the reading on scale D.

[3] It is realized that several statistical technicalities are involved in this statement; however, they are not of practical significance, in view of the fact that the stated probability is not numerically exact.

the lot favorably, as explained in Chapter I.) Appropriate correction factors to allow for the effect of sample size in estimating the true standard deviation can be readily calculated from Student's [4] distribution of σ, and tables of these factors are published. However, the observed standard deviation based on a sample of size n having been calculated, the Alignment Chart for Reading the Estimated Lot Standard Deviation will yield at once the value corrected for the bias of sample size. Thus, by placing a straight edge on the chart and lining up 20 (the sample size) on scale E with 19 (the significant digits of the observed standard deviation) on scale B, one reads 20 (the estimated lot standard deviation) on scale D. Since the correction is small, the position of the decimal point is obvious, and the estimated lot standard deviation is, of course, 0.020.

Estimation of the lot standard deviation. However, this estimated lot standard deviation, like all statistics based on observations, may be greater or less than the true standard deviation, owing to the effect of chance in the selection of samples. The estimate is based on a sample of only 20. One wishes to know the probable limits within which the true standard deviation lies. The same chart supplies these limits. By lining up 20 (the sample size) on scale A with 19 (the significant digits of the observed standard deviation) on scale B, one reads 62 and 93 on scales C and D, respectively. Placing the decimal point by inspection,[5] we may therefore say that it is fairly probable (probability of the order of 0.75 to 0.90) that the true lot standard deviation lies between 0.020 ± 0.006, and it is highly probable [6] (probability of the order of 0.89 to 0.95) that the true lot standard deviation lies between 0.020 ± 0.009. One must ask oneself, is this close enough? Only engineering judgment can yield a practical answer, and all the statistics under the sun cannot relieve one of this burden. If the answer is no, then one must take more

[4] *Biometrika*, Vol. 6, 1908, pp. 1–25. Previously published by F. R. Helmert, *Astronomische Nachrichten* 88, No. 2096, 122 (1876). See footnote on page 133.

[5] If n is between 4 and 100, the reading on scale D is less than the observed standard deviation and lies between σ and 0.2σ. If n is less than 5, the reading on scale D is slightly greater than σ. No other rules for pointing off are necessary. The reading on scale C is always $\frac{2}{3}$ the reading on scale D.

[6] One may point out that this is an inverse probability—so it is. However, the avoidance of the inverse probability generally has little effect on numerical results, and the statement is valid because the stated probability is not numerically exact but is merely of the order of thus-and-so. It is believed that practical application would merely be rendered more complicated and not improved by more refined procedure, and this simplifying but not misleading approximation will be found in subsequent instances of this character.

samples. If one wished to get the so-called probable error (theoretically 0.5 probability), it could be readily calculated by multiplying the D-scale reading by 0.2248, which yields an answer of $\pm$ 0.002 in the present case. However, probable error is rapidly falling into disuse; hence no scale is given for reading this corresponding probability.

Thus, given the 20 observations regarding the nominal 2-inch lengths of rod, one would estimate the average length of the lot to be 1.97 inches and judge that the individual lengths are distributed about the average length with an estimated standard deviation of 0.020 inch; and one would make this judgment well knowing that these figures which are based on a sample of 20 are inexact as indicated by the probable limits. The important thing about knowing the average and standard deviation consists in being able to judge therefrom the distribution of the whole lot of articles with respect to the quality characteristic under investigation; in this case, length. Since plus and minus three standard deviations includes practically the whole universe, one would judge that, if one continued to cut nominal 2-inch lengths in this same way, practically all lengths would lie between 1.97 inches $\pm$ 3 $\times$ 0.02 inch, or between 1.91 inches and 2.03 inches. It should be noted that these calculated limits are much wider than the limits between which the 20 observations actually fell, viz., 1.95 inches to 2.01 inches inclusive. It is generally the case that, in drawing a small sample from a large lot (an infinite lot, in the example cited), the extreme members can scarcely be expected to be included; and the calculated limits for the universe or lot will considerably exceed (and justly so) the observed limits in the sample.[7] However, in order to make this type of generalization, prior knowledge must exist to the effect that this type of product is approximately normally distributed, or else certain types of tests should be made which will be discussed a few paragraphs later.

At this point a moment's digression appears advisable. Three frequency distributions have been discussed, and not just one. The

[7] If one wishes other limits, the following formula is recommended:

$$1 - \frac{1}{2.25t^2},$$

where P is the proportion of the lot of articles and t is the number of standard deviations. Thus, taking $t = 1.5$, one has $P = 0.80$. Approximately 0.80 of the lot of articles are between 1.97 $\pm$ 1.5(0.20) or 1.97 $\pm$ 0.30. This formula is in general somewhat conservative and may tend to promise of slightly less than true proportion between limits. Its nature will be enlarged upon in Chapter X.

first is the frequency distribution of the quality characteristic, X (length in this case), which is represented schematically in Figs. 5·1 and 5·2. The estimated values of the average and standard deviation of this distribution were calculated, and as a consequence of these values one can draw various deductions regarding the frequency distribution; e.g., that it is highly probable that practically all values of X be between 1.91 and 2.01 inches ($\overline{X} \pm 3\sigma$).

The distribution of averages. This frequency distribution gave rise to another. That is to say, if samples of 20 are drawn from the parent distribution, and the average observed, some averages will be greater than others. These averages also form a frequency distribution. This frequency distribution also has an average and a standard deviation. The average and standard deviation of this frequency distribution bears a statistical relation to the parent distribution, and could be calculated; and as a result of these calculations one could draw various deductions about the frequency distribution of the average of samples of 20; e.g., that it is extremely probable that the true average lies between 1.97 ± 0.014. In fact, such calculation is the normal process; however the chart for means did this work. We merely put certain statistics of the parent distribution in it; turned the crank, so to speak, and it ground out the probable deductions.

The distribution of standard deviations.[8] In like manner, the parent frequency distribution gave rise to a frequency distribution of the standard deviations of samples of 20, which itself possessed an average and standard deviation. It was not necessary, however, to calculate these statistics (the average and the standard deviation), as the chart for standard deviations gave the kind of probable judgment needed for practical use.

It is not proposed to burden the reader with a theoretical discussion of statistics; but if the concepts of the frequency distribution of the statistics (average, standard deviation, etc.) of a parent frequency distribution are kept in mind, the philosophy and logic underlying all future discussion of statistical methods of dealing with variables will become most agreeably simple.

The test of data for validity of predictions. The procedure outlined thus far will not appear very complicated even to one who has never seen it before. Its application is exceedingly quick and simple,

[8] For an excellent discussion of the subject matter implied by this chapter, see *On the Statistical Theory of Errors*, W. Edwards Deming and Raymond T. Birge, The Graduate School, Department of Agriculture, Washington, D. C., 1938.

and indeed the procedure has been applied in one form or another for many years. However, there is a fly in the ointment. Except for admitting that the probability is inexact, the principles governing the procedure outlined are much the same as those on which time-honored P.E. is based. The question was not raised whether one appeared to be justified in applying this procedure to these data. This is a very grave question, and it was never seriously raised from the time of Gauss until the very recent present. This inaction was due largely to ignorance, but is also attributable to the fact that no practical way was known of getting an answer to the question. The failure to raise this question has at times led to very distressing results. Shewhart [9] has shown that, even in such precise work as Michelson's determinations of the wave length of light and Heyl's determination of the acceleration due to gravity, an application of this procedure for determining probable limits yields results which are contradicted by subsequent and equally precise determinations. That is to say, quoted measures of precision, if quoted at all, were frequently wrong. Why? Because in applying this procedure one tacitly assumes that the data behave as random drawings from a statistically uniform product whereas the data may be (and not infrequently are) subject to assignable causes of variability. If one is making predictions on which no one can or will check, perhaps he can ignore this point with impunity; but if one is rendering the predictions as a guide to practical operational use, as is generally the case with the industrial statistician, hundreds or thousands of articles will be checked in accordance with his prediction, and he is almost certain to be caught if wrong. Hence, it appears highly advisable to make a check on the statistical uniformity of the data—i.e., absence of assignable causes of variability—for, unless the data are statistically uniform (and this means the data of the universe), one simply cannot reliably predict from sample to lot. Fortunately, modern statistical methods offer a check which is indeed simple, and it will be illustrated with the data used in the example just cited.

Illustration of test for assignable causes of variability. Divide the data into small rational subgroups, according to some engineering reason, if such exists; otherwise, into mere arbitrary groups in the order of observation; then take the average and standard deviation of each subgroup. By dividing the 20 observations of length into 5

[9] Chapter III, *Statistical Method from the Viewpoint of Quality Control*, W. A. Shewhart, 1939, The Graduate School, Department of Agriculture, Washington, D. C.

subgroups of 4 observations each in the order of observation, the following statistics result:

$\bar{X}_1 = 1.9775$	$\sigma_1 = 0.0238$
$\bar{X}_2 = 1.9650$	$\sigma_2 = 0.0112$
$\bar{X}_3 = 1.9650$	$\sigma_3 = 0.0300$
$\bar{X}_4 = 1.9550$	$\sigma_4 = 0.0100$
$\bar{X}_5 = 1.9875$	$\sigma_5 = 0.0296$
$\bar{\bar{X}} = 1.9700$	$\bar{\sigma} = 0.0209$

Then treat $\bar{\sigma}$ as an observed or uncorrected standard deviation, and from the alignment charts, obtain the scale-D reading for n = the subgroup size (4 in this example), and $\sigma = \bar{\sigma}$ (0.0209 in this example).

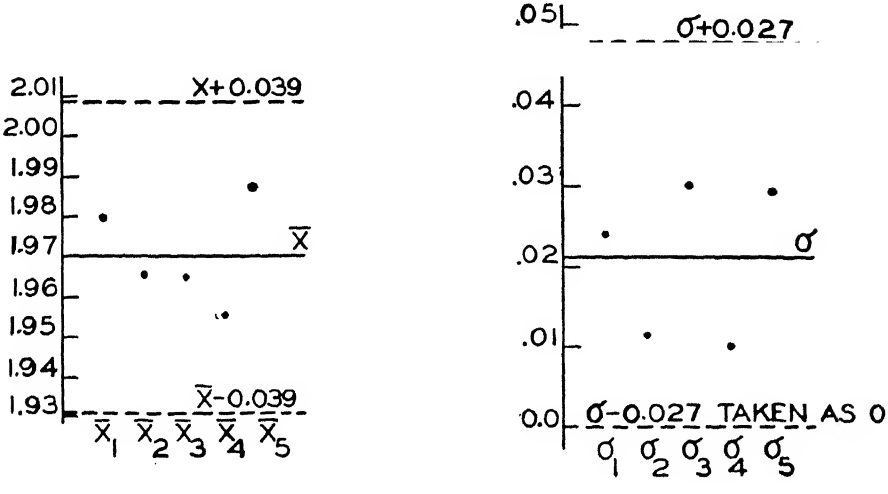


FIG. 5·3. Test for Assignable Causes of Variability by Control Chart Method.

This consists of nothing more or less than solving the frequency distribution of averages for samples of 4 for its probable limits and of solving the frequency distribution of standard deviations of samples of 4 for its probable limits. The reading for averages is $\pm$ 0.039, and the reading for standard deviations is $\pm$ 0.027. If a $\bar{X}$ should fall outside of 1.97 $\pm$ 0.039 or any σ outside of 0.021 $\pm$ 0.027, it is highly improbable that the data are statistically uniform, for, if the subgroups were really taken from a statistically uniform parent lot, the probability is very high against this occurrence. A simple control chart plot is strongly advised in this case (see Fig. 5·3), as the graphic exposition has many virtues. If any points should fall out-

side the control chart limits, one can quite assuredly conclude that no reliable prediction is possible.

No points fall outside the control chart limits; hence there is no reason for not believing that the data are statistically uniform. It does not follow that the universe is necessarily statistically uniform; it merely has not been proved otherwise. However, experience shows [10] that the meeting of this criterion by 25 subgroups of 4 is indeed reliable assurance of uniformity. It is sometimes necessary to accept the hazard of making tentative predictions, based on satisfaction of the criterion by a considerably smaller number of samples of 4, when the sampling is expensive or further sampling impossible and *there are cogent engineering reasons for believing the product to be uniform.* Under such circumstances the probable inference, of course, becomes more a matter of engineering judgment and less a matter of statistical analysis. However, in offering tentative predictions based on scant data one must recall the story of the weary man who went into the restaurant and ordered coffee, toast, and a soft-boiled egg. "Would you care for anything else?" asked the waitress.

"Yes," said the customer, "A few kind words." As the waitress started silently away, after bringing his order, he asked, "How about the few kind words?" She deftly whispered, "Don't eat the egg."

[10] *1933 A.S.T.M. Manual on Presentation of Data*, Second Printing, March 1937, par. 7.

CHAPTER VI

INSPECTION OF A NUMBER OF RELATED LOTS OF ARTICLES—SAMPLING BY VARIABLES

Alle Kunst ist umsumst
Wenn ein Engel auf das Zündloch brunzt.
—*An Old Folk Saying*

Comparison of the grand-lot systems for attributes and variables. The discussion of the principles underlying the grand-lot scheme for exploiting engineering knowledge in Chapter III and the discussion of the nature of variables in Chapter V have laid a foundation for the development of a grand-lot scheme for the inspection of related lots when sampling by variables.

It is presumed in this case that (*a*) one is sampling lots of articles which one has had on hand for some time or (*b*) that one is sampling consignments from some company or manufacturer. The control of quality of one's own product is the subject matter of Chapter VII. Of course, it is further presumed that one inspects the lot with a view to estimating the constants of the distribution (in this case the average and standard deviation), in order that by means of these constants one can further estimate the limits within which practically all the articles of any lot will lie, or the limits within which any assigned proportion will lie. The means of making such estimates, knowing $\overline{X}$ and σ, were covered in Chapter V.

The grand-lot system devised for attributes cannot merely be extended to apply to variables, because of some fundamental differences in the two problems. In the interest of clarity it appears advisable to consider some of these points of difference at the outset.

Variables are inherently more complicated than attributes. In dealing with attributes there is no question about the functional form of the universe (whether it is bell shaped, skewed, flattened, etc.) and there is but one parameter of interest, viz., the fraction defective. One cannot know the exact grand-lot fraction defective from the observation of samples, but one can know it very closely; and having chosen a grand-lot fraction defective, one can, subject to the sole

assumption of randomness of sample, calculate absolutely rigorous probable limits within which the observed sample fraction defective should lie.

In dealing with variables, it almost appears that chance has systematically introduced obstacles to hinder one who would scientifically gain probable knowledge. The functional form of the universe is practically never known, and to follow classical error and assume it to be normal would surely be dangerous in practical work. One must deal with at least two parameters (the average and the standard deviation), and, up to the present time, theoretical knowledge regarding the distributions of these parameters is incomplete. Hence, even with values of the average and standard deviation assigned, exact probable limits cannot be calculated within which sampling results should lie; and anyone who claims so to calculate them is either misinformed or indulging in such looseness of statement as to distort facts misleadingly. However, as was indicated in the preceding chapter, sufficient knowledge appears to be available for calculating limiting probabilities; e.g., the probability is *at least* 0.95, 0.90, etc., that a statistic lies between certain limits.

By way of advantage, however, sampling of variables is, in general, much more efficient with respect to sample size than sampling of attributes; and consequently required sample sizes are in general much smaller.

Pursuant to these observations, grand-lot schemes can be set up for sampling by variables. Of course, there should at least be engineering reasons for believing (*a*) that the lots of the grand lot are essentially the same, and (*b*) that the respective lots were made under conditions purporting to insure uniformity. Furthermore, two schemes must be applied, one for averages and one for standard deviations; also, both schemes must be applied to each quality characteristic one chooses to investigate.

The grand-lot scheme for standard deviations. Suppose that one has 15 lots of articles which on engineering grounds are believed to be essentially the same with respect to the quality characteristic, X, which is sampled by variables. The matter of sample size will be covered in Chapter X, but for the present let it merely be stated that the grand-lot sample size should be relatively large, and that the lot sample size should be at least 5. Suppose that samples of 5 are taken from each of the 15 lots, and the average and standard deviation of the sample from each lot calculated with results as shown in the

second and third columns of Table 6·1. The table also shows the detailed computation of the average and standard deviation of lot 1, the average of the averages, $\bar{\bar{X}}$, and the average standard deviation, $\bar{\sigma}$.

TABLE 6·1

STATISTICS OF A GRAND LOT OF ARTICLES

Lot	$\bar{X}$	σ	n	Lot	$\bar{X}$	σ	n
1	10.14	0.15	5	9	9.95	0.14	5
2	10.02	0.06	5	10	9.93	0.12	5
3	9.80	0.18	5	11	10.05	0.12	5
4	9.90	0.34	5	12	9.90	0.22	5
5	9.98	0.19	5	13	9.66	0.45	5
6	9.82	0.17	5	14	9.84	0.14	5
7	9.88	0.18	5	15	9.82	0.21	5
8	10.01	0.05	5				
					$\Sigma\bar{X} = 148.70$	$\Sigma\sigma = 2.72$	
					$\bar{\bar{X}} = 9.91$	$\bar{\sigma} = 0.181$	

LOT 1

X	$(X - \bar{X})$	$(X - \bar{X})^2$
10.24	0.10	0.0100
9.96	0.18	0.0324
10.26	0.12	0.0144
10.29	0.15	0.0225
9.96	0.18	0.0324
50.71		0.1117

$$\bar{X} = 10.14$$

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}}$$

$$= \sqrt{\frac{0.1117}{5}} = 0.15$$

It was observed in the previous chapter that calculations regarding either the average or the standard deviation involve the true but unknown standard deviation of the universe. In this case the universe is the grand lot, and the samples from the lot are merely regarded as random samples from a subgroup of the grand lot. One therefore wishes an estimate of σ', the true grand-lot standard devia-

tion. This could be obtained at once from the Chart for Standard Deviations by setting off 5 on the E scale, 0.181 (which is $\bar{\sigma}$) on the B scale, and reading 0.215 (which is the estimated σ') on the D scale, if one only knew that the grand-lot judgment satisfied certain criteria. It therefore appears that one should first test the grand-lot judgment with respect to dispersion. This is easily accomplished by means of a grand-lot scheme for standard deviations.

The first thing one wishes then is a pair of inner limits, beyond which any point is a suspected maverick; and a pair of outer limits,

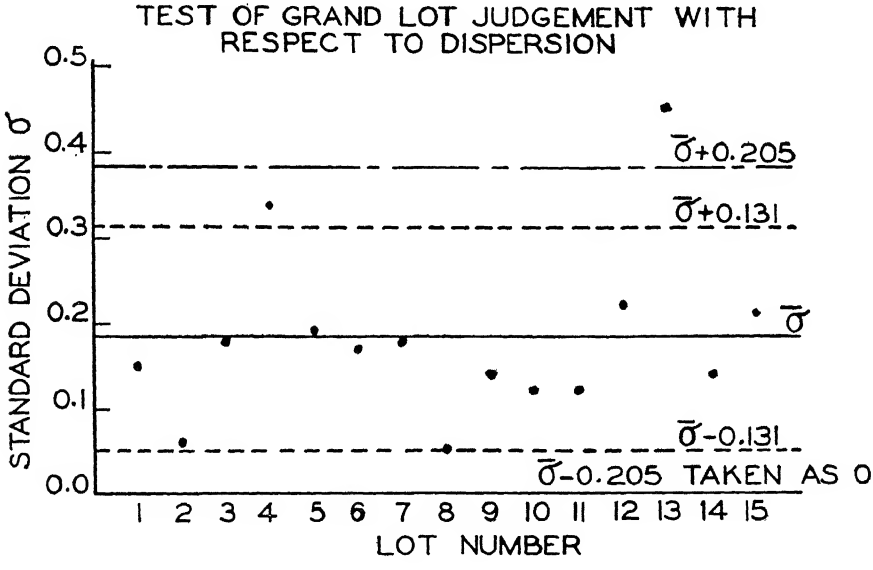


FIG. 6·1.

beyond which any point necessitates recalculation of the average statistic (in this case $\bar{\sigma}$). Entering the Chart for Standard Deviations with 5 (the lot sample size) on the A scale, and 0.181 (the average statistic) on the B scale, one reads[1] 0.131 on the C scale and 0.205 on the D scale. Limits are calculated as 0.181 ± 0.131 (for suspected mavericks) and as 0.181 ± 0.205 (for extreme suspected mavericks). This amounts merely to saying that a lot which produces an average standard deviation of 0.181 for samples of 5 has a fair probability (of the order of 0.80) of producing a standard deviation of between 0.181 ± 0.131 for any sample of 5; and a high probability of produc-

[1] Pointing off by inspection as explained in Chapter V.

ing a standard deviation between 0.181 ± 0.205 (lower limit taken as zero).[2] Figure 6·1 is a plot of this analysis.

One need go no further. Lot 13 is an extreme suspected maverick, and must be eliminated from computations. Its σ is 0.45; therefore, subtracting 0.45 from $\Sigma\sigma$, which is 2.72, and dividing by the reduced number of lots (14), one has a new $\bar{\sigma}$ of 0.162. Entering the Chart for Standard Deviations with $n = 5$ on the A scale, $\sigma = 0.162$ on the B scale, one obtains new probable limits of 0.162 ± 0.122 and 0.162 ± 0.183. Figure 6·2 is a plot of this analysis.

For the sake of clarity, one should note just what this figure means. It merely states that, if all the samples (except lot 13) had been drawn from a single grand lot of average standard deviation

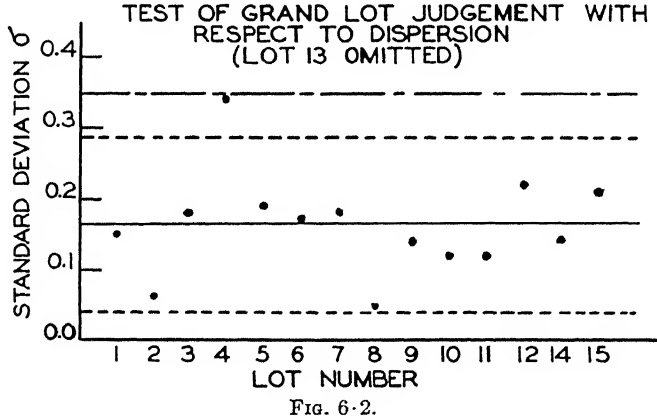


FIG. 6·2.

0.162 for samples of 5, it is fairly probable that the standard deviation of any sample of 5 would fall between the inner limits and highly probable that it would fall between the outer limits. This is almost precisely the situation encountered in the grand-lot scheme for sampling by attributes, and the inner limits are at sensibly the same probability. Lot 4 is then, under this criterion, a suspected maverick.[3] The test for rejection of the grand-lot judgment is obviously

[2] The limits for suspected mavericks are merely ± 2 standard deviations of the standard deviation; the limits for extreme suspected mavericks are ± 3 standard deviations of the standard deviation. The order of probability quoted for these limits is much lower than that associated with a like number of standard deviations of a statistic under the familiar normal law because of various unknowns connected with the distributions involved. See Chapter X.

[3] As in dealing with attributes, we abandon the engineering hypothesis of essential sameness of lots, when the observed sampling results are such that the probability is of the order of 10 to 1 against their occurring under the conditions of the hypothesis.

the same as in the case of attributes; i.e., the probability is approximately 0.1 that a bona fide member of the grand lot will be a suspected maverick above the line.[4] Looking up $n = 14$, $Q = 0.1$ on Chart 0.1 = I_Q, one reads a c of between 2 and 3; i.e., the probability is approximately 0.9 that not more than 2 suspected mavericks will occur above the line in a bona fide grand lot of 14 member lots. Since there is only 1 suspected maverick, there is no occasion to reject the grand-lot judgment. Therefore, all lots except 4 and 13 should be assigned the grand-lot standard deviation, for one freely admits that, on the basis of a sample of only 5, he cannot distinguish between the other lots. The best estimate of the grand-lot standard deviation is that σ which would produce an average standard deviation of 0.162 in samples of 5. By entering the Chart for Standard Deviations with 5 on the E scale and 0.162 on the B scale, one reads the estimated grand-lot σ as 0.193 on the D scale. Whether one further tests lots 4 and 13 or otherwise disposes of them is dependent upon practical considerations.

The grand-lot scheme for averages. In addition to knowing the standard deviations of the lots one must also know the averages of the lots, before one can judge even approximately the nature and extent of the frequency distribution[5] of the quality characteristic. A grand-lot scheme for averages is therefore essential. Even to a reader with no previous statistical knowledge, the design of such a scheme is, by this time, rather obvious.

The first question that arises is whether lots which were suspected mavericks or extreme suspected mavericks with regard to dispersion should be eliminated from consideration with regard to averages. It is obvious that a lot could be of quite acceptable average level and still have a very large dispersion. Nevertheless, the standard deviation is an essential parameter in the distribution of the average; and consequently not so much precision could be expected of the averages of 5 taken from lots of apparently greater standard deviation as from those of apparently lesser standard deviation. The practical answer is more a matter of engineering than of statistics. However,

[4] For both practical and statistical reasons it appears better practice to count only suspected mavericks above the inner limit.

[5] Both average and dispersion are not always of great interest. For example, in many types of ammunition a considerable latitude can be allowed in the average, because appropriate correction can be set on the sights; but dispersion retains its practical importance. One's major interest is then in the frequency distribution of the standard deviation rather than the parent distribution of the quality characteristic.

a moment's reflection will show that little harm can be done by retaining the suspected mavericks with regard to dispersion in the grand-lot scheme for averages, but little could be gained by retaining the extreme suspected mavericks, as one would be suspicious of them whatever the result. This policy is therefore recommended.

If, therefore, referring to Table 6·1, the X of lot 13, which is 9.66, is subtracted from $\Sigma\bar{X}$, which is 148.70, and the difference is

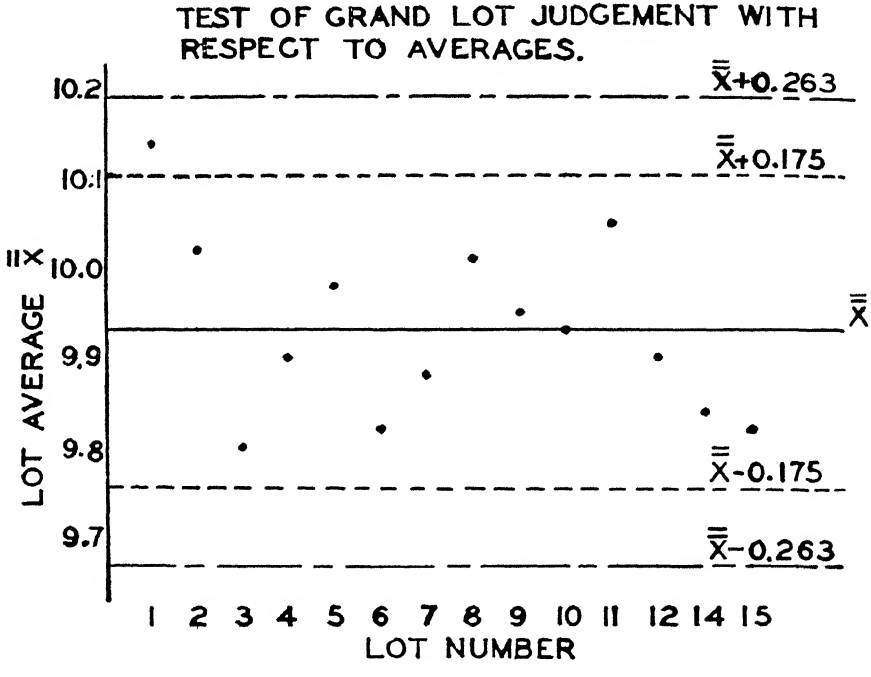


FIG. 6·3.

divided by 14, one gets a new $\bar{\bar{X}}$ of 9.93. Entering the Chart for Averages with $n = 5$ on the A scale, $\sigma = 0.162$ (the $\bar{\sigma}$ of the accepted grand lot), one reads[6] the inner and outer limits[7] on the C and D scales as $\pm$ 0.175 and $\pm$ 0.263. Drawing limits at 9.93 $\pm$ 0.175 and 9.93 $\pm$ 0.263 and plotting the points results in Fig. 6·3. There are no extreme mavericks, so the computation of the estimate of the grand-lot average, $\bar{\bar{X}}$, need not be repeated (the average of the

[6] Pointing off by inspection, as explained in Chapter V.

[7] The inner and outer limits are at ± 2 and ± 3 standard deviations of the average, respectively.

observed averages is the estimated average; and the average of the observed standard deviations must be corrected for sample size by means of the chart, in order to get the estimated standard deviation). Lot 1 is the only suspected maverick.

Now let one observe that the limits in the case of averages are placed at a higher probability than in other cases.[8] The probability is approximately 0.1 that a lot of true average equal to the grand-lot average will yield a suspected maverick; therefore the test for rejection of the grand-lot judgment is different in the case of the scheme for averages. As before, one enters Chart 0.1 = I_Q with n equal to the number of lots and Q equal[9] to 0.1, and reads the corresponding c;

TABLE 6·2

FINAL GRADES OF LOTS OF A GRAND LOT

Lot	Estimated $\bar{X}$	Estimated σ	Lot	Estimated $\bar{X}$	Estimated σ
1	Unknown	0.193	9	9.93	0.193
2	9.93	0.193	10	9.93	0.193
3	9.93	0.193	11	9.93	0.193
4	9.93 *	Unknown	12	9.93	0.193
5	9.93	0.193	13	Unknown	Unknown
6	9.93	0.193	14	9.93	0.193
7	9.93	0.193	15	9.93	0.193
8	9.93	0.193			

* An apparent figure, but not worthy of quite so much weight as the other tabulated figures.

but in the case of averages the total of the suspected mavericks both above and below the limits must not exceed c. In this example c is 2 and the total of the suspected mavericks both above and below the lines is only 1, so that the grand-lot judgment is not rejected.

The complete results of this inspection are given in Table 6·2. It is worth while to note that the unknown $\bar{X}$ of lot 1 casts no doubt on its σ of 0.193, because its variation was determined about its own observed average and its σ is on an equal footing with the other σ's. However, the unknown standard deviation of lot 4 casts doubt on

[8] Reasons for the shift to higher probability can be inferred in part from Chapter V and in part from Chapter X.

[9] Grounds can be given for choosing Q somewhat smaller, and within the range 0.05 to 0.1.

its average of 9.93 because a greater dispersion in this lot might account for its observed average having by chance fallen between the probable limits when its true average was somewhat distant from the grand-lot average. That is to say, the scheme for averages is dependent upon true standard deviations; the scheme standard deviations is independent of true averages. It might also be noted that lot 13 would have been a suspected maverick if it had been included in the grand lot for averages, and the fact that it was a suspected maverick below the bottom line would not even lend much assurance that its true average was not greater than the grand-lot average; for, with σ unknown, anything might happen.

Results of the inspection system. One may well ask what has been gained by this inspection system? Or, more specifically, how much better is Table 6·2 than Table 6·1? One might answer that it is better by the value of the engineering judgment that went into the system, but that would be in the nature of an understatement, for the system minimized human bias and frailty, and applied the engineering judgment systematically, so as to admit of reproducibility and promote consistency. One does not need an expert to render these grand-lot judgments. A clerk can do it on a basis of simple rules, such as lots made by the same manufacturer, under the same drawings and specifications, at about the same time, etc. In fact, a card file system of data, some rules, a clerk, and the occasional supervision of a qualified engineer is likely to prove better than some good old-time experts. Under Table 6·1 there would have been 15 judgments of standard deviation ranging from 0.05 to 0.45 and 15 judgments of $\overline{X}$ ranging from 9.66 to 10.14, neglecting corrections for sample size which would not improve their consistency. This is an overwhelming percentage variation. Under Table 6·2 one frankly admits that 4 of these 30 statistics are unknown, for judgment was at variance with the evidence of the sample, in these instances. It is worth quite a bit to know how much one does not know. As for the remainder of the statistics, one frankly admits that the $\overline{X}$'s are not exactly the same and the σ's are not exactly the same; but one also knows that, based on the sample size, they appear to be so near to the same that he cannot detect a significant difference between them. That is to say, statistically one lot represented by these values is the same as another, and this statistical judgment checks with the engineering judgment.

An estimate can be made, however, of how much a member of the grand lot can be expected to differ from its assigned value. If a lot is

of true standard deviation which would correspond to a central line on Fig. 6·2, the probability of its inclusion in the grand lot is approximately 0.8; if it is of true standard deviation corresponding to either of the dotted lines marking the inner limits, its probability of inclusion in the grand lot is 0.5; if it is more remote, the probability of its inclusion is, of course, less than 0.5. Since the central line is predicated on a mean, and since the grand-lot judgment that the lots are

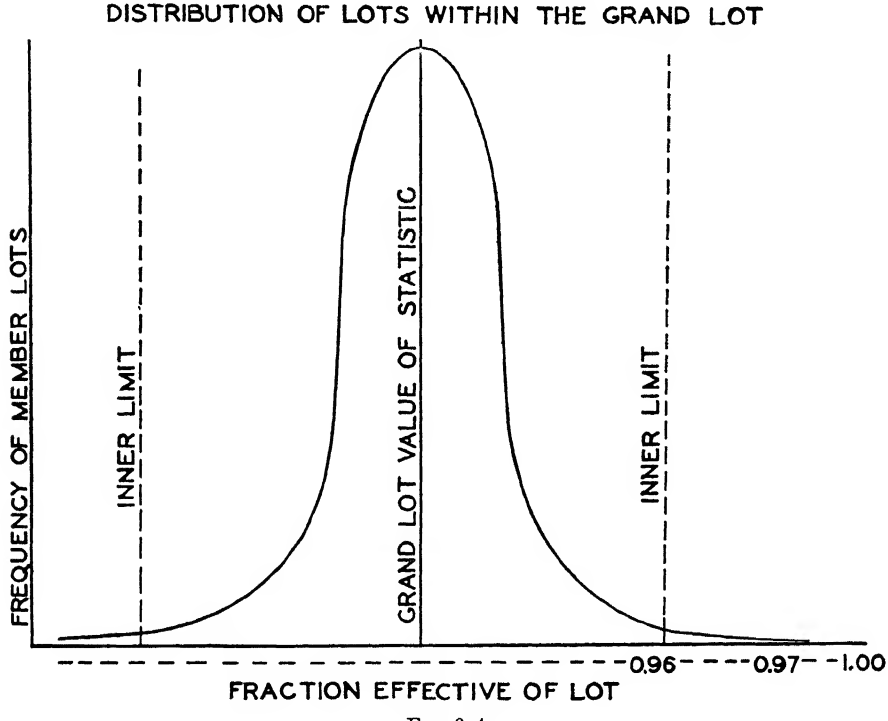


FIG. 6.4.

essentially the same should be of at least some value, it stands to reason that the lots of the grand lot should be considerably more closely clustered about the mean value than at the limiting values. Hence, the probability is very small that a final grand lot will contain any considerable number of lots whose true standard deviations differ from the grand-lot standard deviation by more than $\frac{1}{2}$ the distance between the inner limits. Consequently one may regard the inner limits as a grand-lot band, in view of the facts that this band should contain practically all members of the grand lot and that a member of

the grand lot may lie anywhere within the band. Of course, the members should be more dense in the center portion. The expected distribution can be represented schematically in Fig. 6·4. The width of the grand-lot band can be reduced as much as one chooses by increasing the lot sample size.

CHAPTER VII

PROCESS INSPECTION—SAMPLING BY VARIABLES

Quoth she, I've heard old cunning stagers
Say fools for arguments use wagers.
—SAMUEL BUTLER: *Hudibras*

Some advantages of quality control. The type of process inspection about to be briefly described pertains to that branch of industrial statistics known as quality control. It is presumed that the reader has some knowledge of the advantages of this powerful technique. As an inspection process it renders an authentic guarantee of the quality of the product, and it promptly detects variations in quality. As an inspection system it is a most efficient method of inspection at extremely low cost, for in general it not only reduces wastage in product due to rejection but also effects a great saving in the labor of inspection. However, its economic advantages extend far beyond this phase, for through quality control one can maintain the variability of the product between limits which are economic: i.e., it gives one a very definite lead as to what variation should (from an economic point of view) be left to chance, and what variability should be corrected, thereby leading to the most efficient use of time, labor, and raw materials. Incidentally, these are difficult types of judgments to attain in practice without such an aid. It is a powerful technique for securing uniformity of product. Only bare outlines will be offered here. For more detailed information, the reader is referred to Shewhart's original work,[1] the American Society for Testing Materials,[1] or the British Publication, *Application of Statistical Methods to Industrial Standardization and Quality Control.*[2] The method of process inspection described in this chapter is based on standard deviation, which, of course, requires the squaring of numbers and the taking of square roots. The application is greatly simplified by means of charts. The method of standard deviation is offered because of its broad applicability and high efficiency, i.e., greater information from a given number of samples.

[1] *Op. cit.*

[2] Edited by E. S. Pearson, procurable from the British Standards Institution, London.

Practicability of simple systems of quality control. However, conditions are frequently met in practice where it is more economical to take a few more samples and avoid computational labor. This is especially true, if the difference enables the employment of less highly skilled labor. In that event the simple method offered in Appendix C may be used to advantage; this method is based on range, i.e., the difference between the greatest value and least value in a sample.[3] Furthermore, the system based on range, although completely described in only six pages, includes the necessary details for setting up the system initially. By way of caution, however, it is recommended that the system based on range be checked for a preliminary period against the system based on standard deviation, before depending upon the former alone.

Adoption of quality control without disruption of existing process. Where production is already established, the institution of statistical process inspection will, in general, occasion no appreciable change in the things or quality characteristics for which one inspects. That is to say, one merely continues to inspect for the same things for which he previously inspected. These are generally the quality characteristics upon which tolerances are placed by drawings and specifications. The change lies in the method of recording (by simple control chart), the interpretation of results, the number inspected (generally less), and in the economic results. One must still inspect for each quality characteristic of importance. The control chart will not tell one what is wrong, but it will tell when the trouble occurred and frequently will predict the approach of trouble before its actual occurrence. In order to make the greatest possible use of the method, it is necessary to keep the control charts up to date so that the process will be under continual surveillance. The order of samples is also of great importance. The location of the assignable cause of variation (trouble) and its elimination are engineering problems. These matters are discussed in more detail in Appendix C. The rating and grading of articles on a basis of merit, taking cognizance of all their principal quality characteristics, constitute the subject matter of Chapter IX.

The details of computing the average and the standard deviation,

[3] Range on a sample of 10 may be said to convey approximately the same information as standard deviation on a sample of 9. If sampling is destructive, it may well pay to do a little additional computational labor and use the smaller sample. If sampling means merely taking another measurement, it may be more economical to take the additional sample and save labor in computation.

and of the use of the alignment charts, are given in Chapter V. With this knowledge, and the background of Chapter IV, the details of the quality control technique in sampling by variables can readily be given by two examples.

How the quality control chart works. Figure 7·1 shows a control chart[4] for a quality characteristic of an article not previously subject to statistical quality control. In the case cited, the article was thought to be very carefully controlled, from an engineering point of view. It sometimes happens that a process not subject to statistical control will show control; but processes subject to the highest type of engineering control frequently show assignable causes for variability. The omitted half-hour points on April 1 indicate the usual difficulties encountered in starting production on a new order. During such a period some type of 100% inspection, usually by the go, not-go method, is advisable. The points outside the extremely probable limits (for averages) and highly probable limits (for standard deviations) observed on April 1 and April 2 indicate assignable causes for variation both in the average and in the standard deviation. Corrective measures taken at that time appear to have cured the trouble with regard to dispersion, since no standard deviations fell outside the limits on the next working day (April 5), but the average weight still seemed to vary unduly from hour to hour. Further corrective measures taken on that date were successful in accomplishing statistical control. Only one dot fell outside limits during the period April 5 to April 19 (a part of the chart is omitted). On that date sampling was reduced (in view of the demonstrated state of control) from 5 samples every 30 minutes to 5 samples every hour.

The relationship between the control chart and tolerances. The question naturally arises regarding the relationship between the information given by the control chart and tolerances set by drawings, specifications, or other authority. In the present state of industry, one seldom encounters a tolerance range expressed in terms of the average and standard deviation. Neither engineers nor the consuming public are as yet accustomed to thinking in these convenient and most enlightening terms. Instead, the tolerance range is generally expressed by stating that the product, meaning every article thereof

[4] The method of computing the average ($\bar{X}$) and the standard deviation (σ) and of avoiding computational labor by reading limits directly from the Chart for Averages and the Chart for Standard Deviations is covered in detail in Chapter VI. In process inspection the hypothesis of essential sameness from sample to sample is much stronger than in inspection of related lots, and limits are always read from the D scale of both charts.

(or at least practically every article thereof), shall be within the limits L_1 to L_2. This matters very little to the operator of the con-

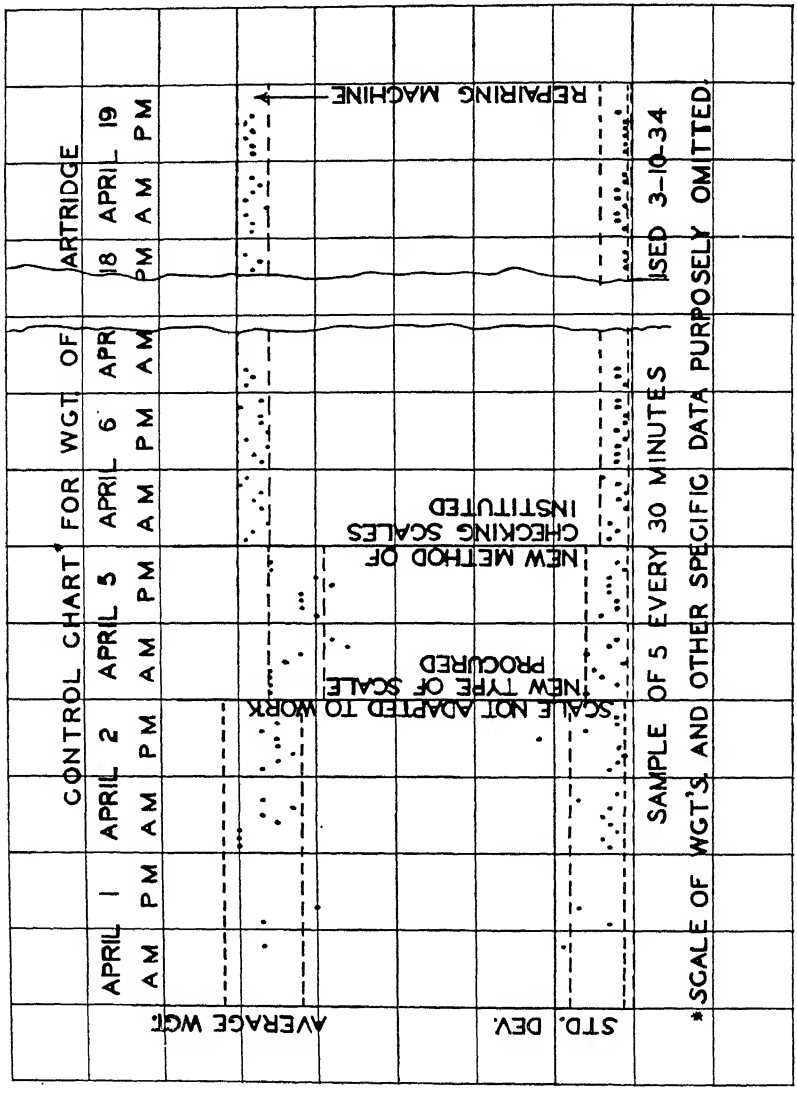


Fig. 7-1.

trol chart. If it is remembered that practically all members of a controlled product will fall within the average plus or minus three standard deviations, it is a very simple matter to inspect the control

chart after the plotting of 25 or more groups of 4 (showing control) for the average and standard deviation and to calculate the limits within which practically all the product will lie. Hence, at least at the end of every lot, a notation should be placed on the control chart in effect as follows:

> "Approximately all products between L_1 and L_2—
> Allowed L_1' to L_2'"

L_1 and L_2 are the over-all limits inherent in the process as calculated from $\bar{\bar{X}} \pm 3\sigma'$ (σ' being obtained from the chart for standard deviations by reading the subgroups sample size on the E scale, $\bar{\sigma}$ on the B scale, and σ' on the D scale). In any sort of quantity production, the number of plotted points will be at least several hundred; and, if the chart does not indicate a lack of control, the manufacturing limits are known with as extreme precision than that associated with the majority of the world's most precise scientific measurements.

The conditions under which percentage inspection is practicable. The relationship between the manufacturing limits L_1 and L_2, and the designated limits L_1' and L_2', is a most important one. If the manufacturing range L_1 to L_2 is definitely less than the tolerance L_1' to L_2', and if there is no difficulty about the location of $\bar{X}'$, percentage inspection with quality control yields most happy results. However, if L_1 to L_2 is greater than L_1' to L_2', that is quite another matter. *Two points deserve utmost emphasis:* (1) *the manufacturing range is inherent in the process, and nothing can be done to reduce it without altering the process;* (2) *if the manufacturing range is greater than the tolerance range, there is no way to meet the tolerance except through 100% inspection.* If the inspection process is destructive, of course 100% inspection is impossible; and inspection, whether or not destructive, is always expensive.

The contribution of quality control to design. It not infrequently happens that upon the introduction of quality control in an established process, the manufacturing range L_1 to L_2 is found to be significantly greater than the tolerance range L_1' to L_2'. This situation arises even when quality control has made a marked improvement in the uniformity of the product. The answer to the paradox is twofold: First, in the absence of statistical methods, the judgment of quality is generally predicated on taking a few small samples and inspecting to see whether or not they meet the criterion L_1' to L_2'.

Such an inspection method does not begin to reveal the total variation in the product, and the interested persons are completely lulled into a false sense of security regarding the uniformity of the product. Second, the tolerance range L_1' to L_2' is often very much smaller than what is actually required (*a*) because experience has appeared to indicate that it had to be reduced and reduced, in order to obtain satisfactory products under the inspection methods extant, or (*b*) because it was set merely by arbitrary authority. Where such a situation is discovered, it is quite in the interest of all concerned to face the facts frankly and revise the standard on a scientific basis.[5]

However, if a tolerance range must be met which is less than the manufacturing range (without 100% inspection and elimination of the non-conforming articles), then the process must be altered so as to yield a product which has a smaller range of chance variation. This is largely an engineering problem. The statistical method merely tells the engineer what he has accomplished. For a statistical method which may help the engineer in this problem, the reader is referred to Chapter XXI of Shewhart.[6]

Some services rendered by the control chart. Figure 7·2 illustrates the beginning of manufacture on an article, where a similar article has been previously manufactured under a system of statistical process inspection. It will be seen that in this instance there is no evidence of lack of control even at the start, and that the sampling is on a very moderate basis. The sampling on which the chart is based was non-destructive and comprised about eight-tenths of 1% of the product.[7]

It should be noted that a file of these charts give a detailed and authentic record of the quality of each lot of product manufactured. The information given is far superior to that which could be given by almost any reasonable acceptance test. The cost of installation of the system may be considerably less than nothing, because under this system of process inspection, it becomes feasible in many instances to replace 100% inspection with percentage inspection with no loss in precision of information or reduction in quality. In fact, quality is almost always improved, wasted labor saved, and cost re-

[5] An actual account of such a procedure is given in "Deviations in Product Prove Machine Performance," L. E. Simon, *Product Engineering*, December, 1936.

[6] *Op. cit.*

[7] As a statistical illustration the chart is as the original, but the data have been altered so as to render it entirely fictitious.

duced. Furthermore, one really knows the quality of the product, which, strange as it may seem, is not always the case even under 100%

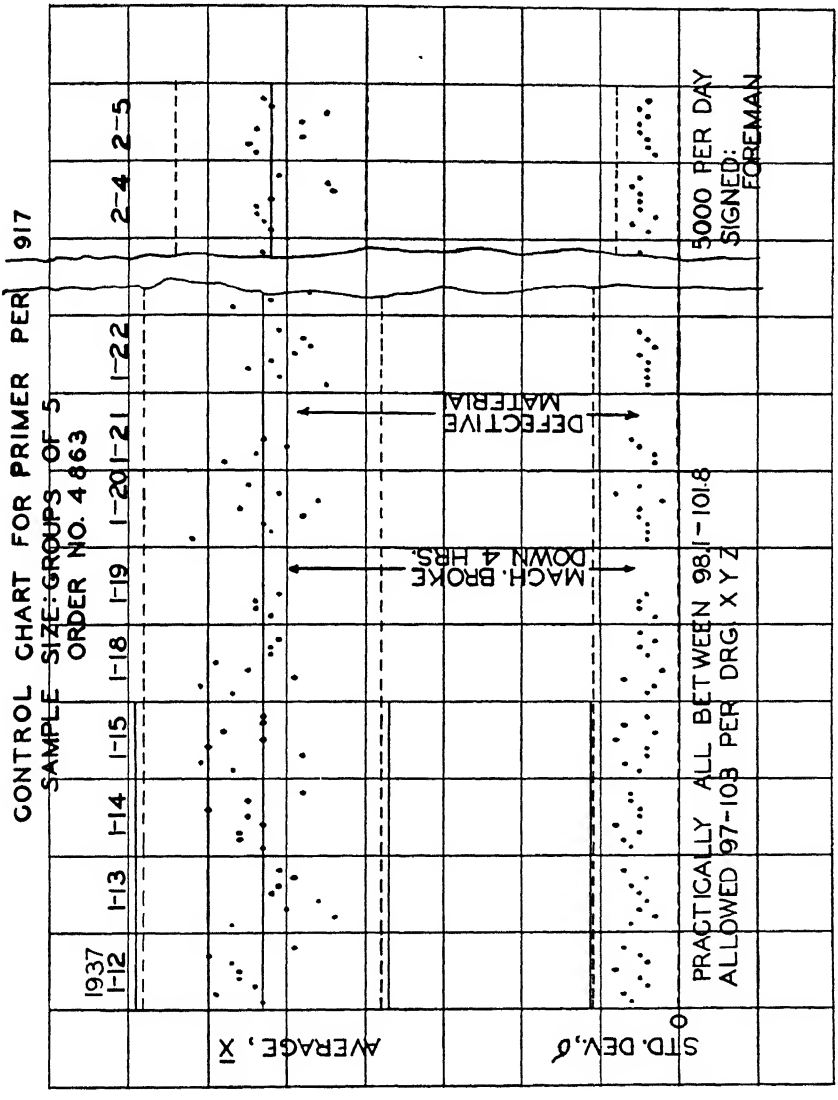


FIG. 7·2.

inspection. Under percentage inspection one can afford much more precise measurements than when all or a large part of the product must be inspected. This type of process control is a potent factor in

making the most economic use of materials, for it effects a factual presentation of the situation with respect to manufacture and inspection; and thereby lays the groundwork for a meeting of the minds between these two divisions and the engineering division; and paves the way for the intelligent establishment of standards which are economic.

CHAPTER VIII

THE SPECIAL CASE OF INDETERMINATE SAMPLE SIZE

All nature is but art, unknown to thee;
All chance, direction, which thou canst not see—
—ALEXANDER POPE: *Essay on Man*

There is a type of sampling which may be regarded as partaking of the nature both of sampling by attributes and of sampling by variables. This type of sampling occurs in the inspection of plated work for surface defects, the examination of insulated conductors on a basis of the number of breakdowns per unit length, the inspection of a surface like photographic film or sheet metal, or the inspection of any finished product for visual defects. It is concerned with the special case of classification of articles according to the number of defects per article as distinguished from a classification of articles into the two simple categories defective and not-defective.[1] It is believed that clarity will be enhanced in this very interesting case by dealing with process inspection, inspection of a single lot, and inspection of related lots, in a single chapter and in the order stated.

Process inspection. This is a type of sampling which production men appear to regard as quite perplexing. However, the problem of measuring this type of quality and of predicting the probable limits of variation due to chance causes is easily solved by statistical methods.

If one were to regard an article which possessed 1 or more defects as a defective, and if practically all articles possessed at least 1 defect, the fraction defective, Q, would be unity, and the sampling

[1] Specifically, the quality characteristic defect is a discontinuous variable. It is a variable because it can take a number of different values. It is discontinuous because it can have only discrete values. If a discontinuous variable is treated by manifold classification, as is done herewith, the treatment results in a series of *quantitative* attributes as distinguished from the mere *qualitative* attributes discussed in Chapters I, II, III, and IV. Of course, if the treatment is reduced to the ultimate crudity of only two classes, one has, in effect, a qualitative attribute. An enlightening discussion of these classifications is found in Chapters 1, 5, and 6 of *An Introduction to the Theory of Statistics*, Eleventh Edition, G. Udny Yule and M. G. Kendall, Chas. Griffin & Co., Ltd., 1937.

would mean nothing. However, a moment's reflection will show that, even in a sample of 1 article, the sample size with respect to the number of possible defects may be almost infinity; i.e., there is an almost infinite number of points at which a defect could occur. If a considerable number of articles have been inspected and the average number of defects per article is $\bar{c}$, the problem of measuring practically the full range of variation due to chance is reduced to merely finding two numbers, c_U and c_L, such that the probability is 0.995 that not more than c_U defects will be observed in a single sample when the average number is $\bar{c}$; and that the probability is 0.005 that less than c_L will be observed under like conditions. It is presupposed that variation is due to chance in random samples. Failure to meet the criterion, of course, implies other than chance causes of variation. This happens to be a type of problem which is capable of ready and rigorous solution.[2] Solutions for all values of $\bar{c}$ up to 40 are given by Fig. 8·1.

Thus, to operate a quality control system on this type of article, one should take a single sample (an article, a given area of a surface material, a given length of cable, etc.) every hour, every 50 articles produced, or other unit of spacing. After having inspected a considerable number of samples, compute $\bar{c}$ and look up the c_U and c_L associated with this value of $\bar{c}$ on Fig. 8·1. Any sample which has a number of defects not within the range c_U to c_L is an indication of an assignable cause for variability.

For example, suppose that, in the production of brass cartridge cases, approximately every fiftieth case from a production line is inspected for folds in the metal or foreign inclusions (surface defects),[3] with results as shown in Table 8·1 (example is fictitious). On looking up $\bar{c} = 5$ on Fig. 8·1, one finds $c_{L0.005} = 1$ and $c_{U0.995} = 11$. The probability is 0.995 that not more than 11 defects will be observed in a single sample, when the average number of defects

[2] These are the ideal conditions for the Poisson exponential binomial limit, which is otherwise known as the law of small numbers:

$$P(>c-1) = 1 - e^{-\bar{c}}\left(1 + \frac{\bar{c}}{1!} + \frac{\bar{c}^2}{2!} + \cdots + \frac{\bar{c}^{c-1}}{(c-1)!}\right),$$

where $P\ (>c-1) =$ the probability of at least c defects.

[3] Cracks which would cause unserviceability would be sampled on a basis of fraction defective (see first part of chapter). Diameter or length would be sampled in like manner, if a go, not-go gauge were used, but would be sampled by variables (see Chapter VII) if these dimensions were actually measured on a continuous scale.

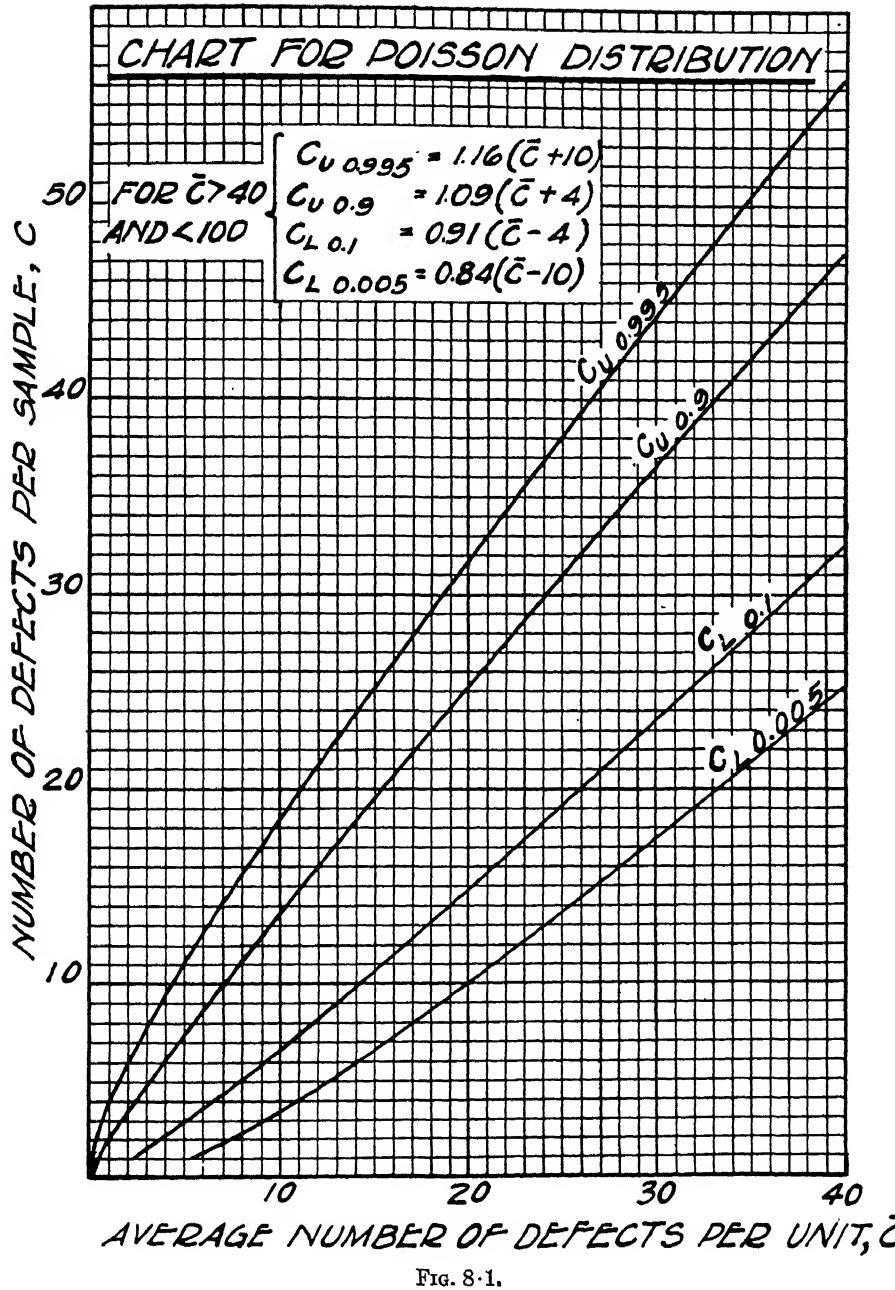


CHART FOR POISSON DISTRIBUTION
FOR $\bar{C}$ > 40 AND < 100
$C_{U\,0.995} = 1.16(\bar{C}+10)$
$C_{U\,0.9} = 1.09(\bar{C}+4)$
$C_{L\,0.1} = 0.91(\bar{C}-4)$
$C_{L\,0.005} = 0.84(\bar{C}-10)$
$C_{U\,0.995}$
$C_{U\,0.9}$
$C_{L\,0.1}$
$C_{L\,0.005}$
NUMBER OF DEFECTS PER SAMPLE, C
AVERAGE NUMBER OF DEFECTS PER UNIT, $\bar{C}$
10
20
30
40
50

FIG. 8·1.

TABLE 8·1

SURFACE DEFECTS IN BRASS CARTRIDGE CASES

Sample Number	Number of Defects	Sample Number	Number of Defects	Sample Number	Number of Defects
1	4	11	2	21	6
2	5	12	5	22	0
3	7	13	7	23	4
4	3	14	4	24	5
5	2	15	1	25	7
6	8	16	4	26	8
7	6	17	2	27	6
8	4	18	7	28	13
9	1	19	5	29	5
10	5	20	9	30	5
				—	—
				Total 30	150
				$\bar{c} = \frac{150}{30} = 5$	

is 5. Figure 8·2 shows a control chart for these data. Sample number 22 and sample number 28 indicate the presence of an assignable

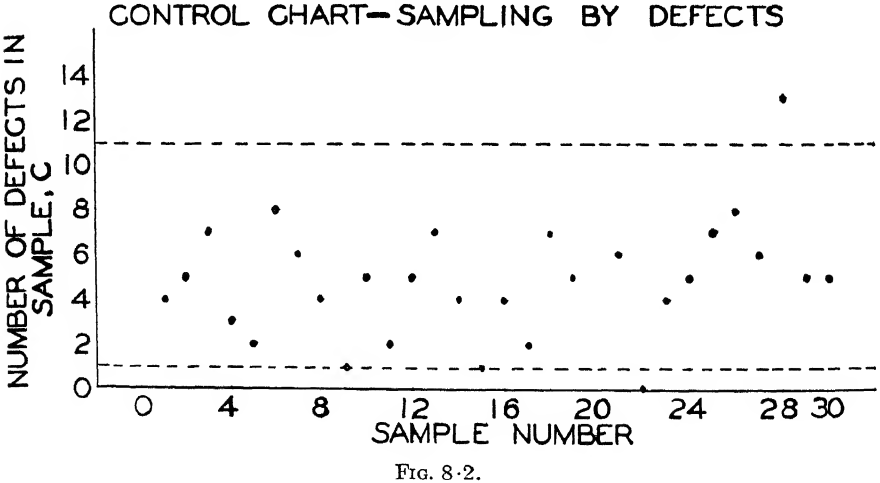


FIG. 8·2.

cause for variability. Steps should be taken to discover the causes of lack of uniformity and eliminate them, for otherwise there is no assurance from taking small samples that some members of the

unsampled portion of the lot do not contain almost any number of defects. That is, without a state of statistical control, one cannot predict lot quality from sample quality.[4]

In a broader sense, analysis by defects is not so much a special kind of classification by attributes as an entirely different consideration. One can readily see that the same data can be analyzed from both points of view, depending upon whether one is interested in defective articles or in number of defects per article. For example, suppose that, in Table 8·1, an article which contains more than 5 defects is not considered fit for service, and hence is classified as defective; and further suppose that the 30 observations be classified

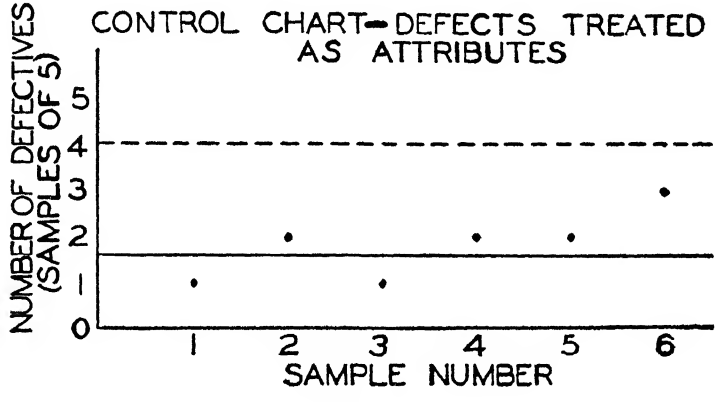


FIG. 8·3.

as 6 samples of 5. The analysis by attributes would then be as shown in Fig. 8·3.

$$\Sigma c = 11, \qquad n = 30, \qquad Q_M = 0.37 \qquad c_L = 0, \qquad c_U = 4.$$

The analysis by attributes is rather ineffective because of the small sample size. However, if sufficient data were given to form, let us say, 10 samples of 40, the analysis by attributes would apply satisfactorily. The two analyses illustrate the importance of clearly under-

[4] At first glance this may appear to be a contradiction of the statements that predictions regarding fractions defective do not require control—only randomness of sample. There is no contradiction. Probabilities regarding the proportions which contain defective articles and the proportions which do not contain defective articles can be calculated irrespective of statistical uniformity. In the case now under consideration the probability has to do with the number of defects in an article, not the proportion of defective articles, and statistical uniformity is essential. The former gives a key only to the quality of the lot as a whole; the test by defects gives a key to the distribution of the quality characteristic within the lot.

standing the thing for which one is analyzing, and the necessity for straightforward thinking in interpreting results. In one case, one analyzed for defects; the sample size was so large as to be almost infinite, even though the sample was one article, and only a few samples sufficed for an informative analysis. In the other case, one analyzed for defectives (a defective being defined as one chose to define it); the sample size was very small, even though the same samples were used as those used in the previous analysis, and change from part to part of the lot (lack of uniformity) could not be detected by such an inadequate sample.

It may be observed from Fig. 8·1 that, after the average number of defects becomes less than approximately 4, one is unable to detect significant improvement in quality, because $c_{L0.005}$ is taken as zero. That is to say, an article which contains zero defects has to be accepted as coming from the same level of quality as articles which contain several defects (significantly poorer quality can, of course, be detected). How can one detect significant variation in quality, when the average number of defects is less than 4? One can accomplish this end by the simple expedient of taking the unit sample as 2 articles instead of 1, and counting the number of defects in the 2 articles. That is to say, by doubling the sample size. In like manner, it can be tripled or quadrupled, if necessary. If one should raise the proposition that the sample of 1 constituted a sample size of the order of infinity with respect to the possible occurrence of defects, and that twice the order of infinity is still the order of infinity, it is suggested that he reflect upon the proposition that all infinities are not equal.

Inspection of a single lot. The reader has probably observed that in this special case, as in sampling of attributes, there is no question of the functional form of the universe. In sampling by attributes, the functional form of the universe was the binomial distribution; here it is the Poisson distribution. However, a particular lot sampled may not be a true lot in the sense that a lot should be composed of members which are essentially alike, but instead may be a pseudo-lot composed of members drawn from two or more universes. As has been previously pointed out, predictions regarding the lot are impossible unless the lot is statistically uniform. A test was provided for uniformity in inspection of a single lot—sampling by variables—but was omitted in the case of attributes, because there are no predictions to be made about this parameter, in the sense

of proportions of a lot within certain limits. Sampling by defects is well suited to a test for uniformity because of the large sample size. Having taken n samples from an unknown lot, and having computed $\bar{c}$, one should then look on Fig. 8·1 for $c_{U0.995}$ and $c_{L0.005}$. The number of defects observed in any sample should fall within these limits; otherwise one should suspect the lot of not being statistically uniform. If the lot is statistically uniform and if $\bar{c}$ is known, then one can predict the limits within which any given percentage of the lot should fall. Two limits are given on Fig. 8·1 and additional ones could be readily computed. Hence, the whole of the information about the lot is given by the one statistic, $\bar{c}$.

Of course, there is a question about how closely one knows $\bar{c}$. This closeness can be judged from the following approximate formulas:

Probability	Limits of $\bar{c}$
0.90	$\bar{c} \pm \frac{1.645\sqrt{\bar{c}}}{\sqrt{n}}$
0.95	$\bar{c} \pm \frac{1.96\sqrt{\bar{c}}}{\sqrt{n}}$
0.997	$\bar{c} \pm \frac{3\sqrt{\bar{c}}}{\sqrt{n}}$

Inspection of related lots. It is quite obvious that, given k lots of articles, one could take n samples from each lot, compute the grand lot $\bar{\bar{c}}$ as $\sum \frac{\bar{c}}{k}$, get the limits for suspected mavericks from the $P = 0.90$ formula, the limits for extreme suspected mavericks from the $P = 0.997$ formula, and thus construct a grand-lot scheme for sampling by defects. However, this type of sampling is in general not destructive, and an inspection of the approximate formulas will show that sample sizes tend to be quite small. Therefore, the sampling of lots individually is not so burdensome as in sampling by attributes; and the grand-lot scheme, although readily available, does not appear to be of great importance.

CHAPTER IX

A METHOD OF EXPRESSING QUALITY [1]

To kerke the narre from God more farre,
Has bene an old-sayd sawe;
And he that strives to touche a starre
Oft stumbles at a strawe.
—EDMUND SPENSER: *The Shepheardes Calendar*

Chapters II to VIII inclusive have given a number of ways of measuring the quality characteristics of lots of articles. These measures, however, are quite unsatisfactory from the viewpoint of expressing the quality of a whole article, principally for two reasons: (1) they bear no clearly defined relationship to any base point as standard; (2) there is no provision for associating the several quality characteristics of an article so as to yield a composite value of the resultant quality arising from the measured quality characteristics. An enumeration of the quality characteristics of an article, each followed by its fraction defective, or its mean and standard deviation, or its mean defects per unit, would indeed present an unreliable basis for judging the relative merit of lots or for judging whether the quality level as a whole was getting better or worse. One could not judge the composite quality as a whole for the mass of detail.

A method will be offered for grading or rating lots or batches of articles on a basis of merit into such grades or ranks (corresponding to good, bad, indifferent, etc.) as one may choose for practical reasons as suitable to his purpose. This method may not apply in all cases, but it has been found applicable and quite satisfactory in practice in a considerable number, and is believed to be of broad application. The importance of some types of quality characteristics must of necessity be assessed by judgment; and consequently the exactitude of the final grade is dependent upon the equitability of these assessments. But the procedure having been established for a type of

[1] An address presented by the author before the Industrial Statistics Conference, Massachusetts Institute of Technology, 1938. Republished with the kind permission of Pitman Publishing Corp., New York.

article, it at least results in consistent judgments of quality, when these judgments are made from time to time, from lot to lot, or by one person or another.

Even the simplest of articles possesses a very large number of quality characteristics. For example, one could trace the quality characteristics of an ordinary commercial caliber 22 cartridge all the way back to the tensile strength, ductility, and hardness of the copper cartridge case; and then on back into the percentage of impurities in the copper, its chemical properties, etc., almost ad infinitum. Happily, however, for practical specification of quality, one need consider only relatively few of the possible quality characteristics. For the caliber 22 cartridge these might well be limited to the muzzle velocity, the firing of the primer (whether it fires or does not fire), and a few observable faults. Therefore, the quality characteristics of an article may well be limited by definition as those quality characteristics which one chooses to consider in evaluating the quality of that article.

Functioning and non-functioning quality. A careful consideration of quality from the point of view of grading lots of articles on a basis of merit leads to the conclusion that two distinct types of quality must be considered. One is functional, i.e., relates to the quality of a lot of articles with respect to the extent to which they accomplish the objective of their design. Quality characteristics significant of functioning are generally capable of rather definite commensuration and are the main subject matter of most specifications. The other aspect of quality is non-functional and pertains to the article in its relation to human wants, desires, prejudices, etc. It is not capable of positive specification, as any number of unforeseen faults may detract from its goodness; and it is generally specified negatively under some such phrase as "any other defects or abnormalities." In order to summarize the quality of articles by rating them in good, bad, indifferent, etc., grades, one must consider both these types of quality for which no common denominator exists;[2] hence one must grade the lots of articles with regard both to functioning quality and with regard to non-functioning quality; and then one may express quality as a whole by quoting both grades, or less precisely by merely quoting the lower of the two grades. For example, if functioning quality were expressed by Grades A, B, C, etc.,

[2] A system which makes use of weighting factors is described in, "A Method of Rating Manufactured Product," H. F. Dodge, *Bell System Technical Journal*, Vol. 7, April, 1928.

and non-functioning quality were expressed by Grades 1, 2, 3, etc., one could express the final quality of a lot as Grade B–3, or, by limiting the expression to the lower grade only, it could be expressed as Grade III.

As a basis of discussion, let it be assumed that three grades (as many could be created as desired) are suitable for our purposes, defined as follows:

First Grade (A or 1)—Good quality, ranging from perfection to the lowest quality one is willing to present under his name.

Second Grade (B or 2)—Indifferent quality, ranging from the bottom of the First Grade to the lowest quality that is suitable for use and which can be marketed under an alternative name.

Third Grade (C or 3)—Bad quality, ranging from the bottom of Second Grade downward; and which must be reworked or scrapped.

The assignment of numerical limits for these grades will be discussed after measures of quality are developed.

In order to measure functioning quality, it appears convenient to classify functioning characteristics in three ways: (1) those which result in success or failure; (2) those which result in success, one or more degrees of partial success, or failure; and (3) those which result in a variable degree of success. These classes may be briefly expressed as measures by (1) failures, (2) partial failures, and (3) variables (of course the first two are attributes).

Measurement of quality by failures. If a lot of articles happens to possess only a functioning characteristic of the first kind, only that fraction of the lot which possesses the functioning characteristic will be effective as successes; hence that fraction (the expected value of which is determined by statistical means) can be called the fraction effective. Manifestly, the fraction effective adequately describes the functioning quality of the lot. If the lot has several such functioning characteristics, it is also evident that the product of the fractions effective describes the functioning quality of the lot, for this fraction describes the ratio of the effectiveness of the lot in question to the effectiveness of a perfect lot. If, for the caliber 22 cartridge, it is determined by the statistical procedures outlined in Chapters II, III, and IV that the fraction effective of the primer is 0.96, then 0.96 describes the effectiveness of the lot with respect to the primer.

Measurement of quality by partial failures. If a lot happens to possess only a functioning characteristic of the second kind, and if it is statistically determined that the fraction A are expected to be

successes, that the fraction a' are expected to be $1/x'$ successes, and that the fraction a are expected to be zero successes (failures), then the fraction $A + (a'/x')$ can be called the equivalent fraction effective, since it describes a lot of like size and effectiveness composed only of successes and failures.[3] If the lot has several such functioning characteristics, the product of the equivalent fractions effective approximately[4] describes the functioning quality of the lot. If in a lot of caliber 22 cartridges it is statistically determined that 0.94 of the powder charges ignite satisfactorily, that 0.04 ignite with a hangfire (perceptible delay), and that 0.02 fail to ignite at all, and if a hangfire is authoritatively accepted as $\frac{1}{2}$ a success,[5] then the equivalent fraction effective of the powder charge is $0.94 + (0.04/2)$ or 0.96.

Measurement of quality by variables. If a lot of articles happens to possess only a functioning characteristic of the third kind, the frequency distribution of this functioning characteristic will be described by a measure of central tendency and a measure of dispersion such as the average and the standard deviation. If it is known or authoritatively accepted that articles possessing values of the functioning characteristic within certain limits are certain fractions of a full success, then, under the assumption of normality,[6] one can readily calculate the fractions of the lot possessing the respective fractions of success. By the preceding paragraph, the sum of the products of the fractions of the lot in each category of success each multiplied by its corresponding fraction of success is the equivalent fraction effective of the lot. If the lot has several such functioning characteristics, the product of the equivalent fractions effective approximately describes the functioning quality of the lot. Thus, in the lot of caliber 22 cartridges, if it were statistically determined by the methods of Chapters V, VI, and VII that the average muzzle velocity was 1110 feet per second and that the standard deviation was 20 feet per second, and if it is authoritatively established that bullets

[3] Obviously A, a', and a are mutually exclusive events each of which is an attribute of the article and each of which could be considered as a functioning characteristic. The artifice of selection and arrangement herein suggested enables one to avoid the complications which would arise from such procedure and to treat the group simply as a single independent functioning characteristic.

[4] An exact solution can be reached by the expansion of products of weighted polynomials.

[5] Actually this type of malfunction is generally associated with the primer rather than the powder charge, and hangfires are regarded more seriously.

[6] A simple method of computation is given in Chapter X.

with muzzle velocities between 1070 and 1130 are successes, those between 1040 and 1070 and between 1130 and 1160 are $\frac{1}{2}$ successes, and those beyond these limits failures, then the equivalent fraction effective of the muzzle velocity is $0.82 + 0.18\ (\frac{1}{2})$ or 0.91.

Functioning effectiveness. It follows from the above that the functioning quality of a lot of articles possessing functioning characteristics of all three kinds is approximately given by the product of all fractions effective and equivalent fractions effective, and that the fraction so obtained describes a lot of like size and effectiveness composed only of successes and failures. Let this product be called the *functioning effectiveness* of the lot. This procedure gives a consistent and rational method of ranking lots, which is not based on a mere arbitrary index but which is actually indicative of the amount of goodness or effectiveness inherent in the lot. For the lot of caliber 22 cartridges the functioning effectiveness is (0.96)(0.96)(0.91) or 0.839. It should be carefully noted that functioning effectiveness is the result of a multiplying process (not an additive or subtracting process), as *a great deal of confusion has been occasioned from time to time by attempts to add failures and partial failures together.*

Non-functioning quality. As previously intimated, non-functioning quality is almost too illusive for specification. It is characterized by the absence of faults. Faults in manufactured articles can be classified in two ways: (*a*) major faults of such serious nature that no articles containing them should ever go to service as they are dangerous to life, property, or the reputation of the product; (*b*) minor faults which are expected to occur at least infrequently even under good manufacturing practice, which have little or no effect on the functioning of the article, are difficult to weigh, predict, or classify, but which nevertheless detract from the desirability of the article. Any major fault should result in a special classification of the lot under the category bad, and relegate it to 100% inspection, reworking, or salvage. But the minor faults, in general, may as well as not be classed all together; and, if this procedure is followed, the frequencies of the respective minor non-functioning faults become merely additive. That is, if by statistical methods A per cent of the lot is expected to have one non-functioning fault, B per cent of the lot is expected to have another non-functioning fault, C per cent, etc., the frequency of non-functioning faults is merely A plus B plus C, etc. It should be noted that non-functioning quality is specified negatively (by absence of faults) and computed by addition; func-

tioning quality is specified positively and computed by multiplication. Thus, if it were statistically determined that 0.01 of the lot of caliber 22 cartridges had the manufacturer's initial poorly stamped thereon, 0.01 of the lot had dented cases, and 0.02 of the lot had corroded cases, then the frequency of non-functioning faults would be 0.01 + 0.01 + 0.02 or 0.04.

Grading of lots. The numerical limits for Grades A, B, C, and 1, 2, 3 previously described are, of course, dependent upon the type of article and must be set in accordance with experience and judgment. Suppose that in the case of the caliber 22 cartridge it had been found from the quality control charts used in process control that this article could be economically and efficiently manufactured with a functioning effectiveness of not less than 0.90 and a frequency of non-functioning faults of not more than 0.05. These values would be likely to be selected as the respective limits of Grade A and Grade 1, respectively. If it were further determined from experience that it was difficult to market the item if its functioning effectiveness were less than 0.80 or its frequency of non-functioning faults greater than 0.30, these would be rather logical limits for Grade B and Grade 2, respectively. Functioning effectiveness below 0.80 would then be Grade C, and frequency of non-functioning faults in excess of 0.30 would be Grade 3. Thus the grading with respect to functioning quality and non-functioning quality becomes a mere comparison of observed quality with defined standards. Under such a grading system, the lot under discussion would obviously be Grade B–1. Of course a grading schedule, a list of functioning faults with penalties, and a list of non-functioning faults must be established as standard for every kind of article subject to the grading system.

It is obvious that the precision of the grading system is dependent both upon the equitability of the penalties assigned and on the choice of quality characteristics selected for consideration. It should also be noted that the quality characteristics should be so chosen as to be independent; i.e., the absence of one characteristic should not influence the presence or absence of another; otherwise the simple procedures outlined are impaired. However, it does yield a common-sense summary of quality which is based on merit, and it has the great advantage of consistency as opposed to the bias and incertitude of casual human judgment.

CHAPTER X

SAMPLE SIZE

Attempt the end, and never stand to doubt;
Nothing's so hard but search will find it out.
—ROBERT HERRICK, *Seek and Find*

The importance of sample size. Sample size is probably of greater interest to the average man than any subject discussed thus far. However, no discussion of sample size is possible without reference to the statistics with which it is concerned. Therefore it appeared necessary to postpone the discussion until the presentation of statistical methods has prepared a minimum background for it.

Even in organizations associated with considerable technical advancement, sample sizes are often set by custom, experience (on which few if any scientific checks have been made), or mere guess. Sample sizes that are set in this manner are sometimes uneconomic, because they are larger than necessary. However, it appears that more frequently they are set too small. In this event, they are likely to be still more uneconomic, for, if they were large, wastage was at least limited to the excess samples, whereas, if they are too small, the paucity of sample may lead to inappropriate action regarding the entire lot. It is obvious, therefore, that there must be an economically right sample size for each sampling problem. The determination of the most economic sample size may, however, be impracticable, but it appears that whoever is connected with sampling should at least know enough of the fundamental facts which bear on sample size to make his sampling procedure logically consistent with his engineering and economic aims and purposes.

Prior essentials for estimating sample size. A moment's digression to get the subject straight will not be wasted. One often hears the question, "What size of sample should I take?" In a logical sense, that is not a question; it is an exclamation, because it is incomplete. The appropriate rejoinder would be, "What size of sample should you take to do what?" and the answer to that "what" leads directly to the fundamental considerations that often escape one's notice.

The nature of these considerations can at least be indicated if the question is reworded as follows: "What size of sample shall I take in order that the probability will be thus-and-so that the error in my estimate of the statistic (average, standard deviation, fraction effective, etc.) will not exceed E?" That improves the matter, but all necessary considerations have not been enumerated, for surely the choice of the probability and of the error limit, E, involve economic considerations; and as indicated in Chapter V some knowledge of the shape of the distribution of the statistic and some knowledge of its dispersion is also essential for calculating an answer. However, in the light of modern statistical methods answers to the question of sample size which are sufficiently precise for practical purposes can be offered.[1]

The requirement of prior knowledge of the fraction effective (in sampling of attributes) and of the dispersion and shape of the distribution (in sampling of variables) seems a bit paradoxical, since, if these parameters were known, the need for sampling would often cease to exist. However, on second thought the requirement is not too onerous, since in practice one nearly always has a sufficiently definite idea of the general magnitude of the parameters or at least of the magnitude that he is interested in detecting to serve as a reasonable datum on which to predicate sample size.

Information gained from sampling would be of academic interest only, unless some action were going to be taken in event the universe sampled appeared to be worse than some predetermined value. Therefore, the mere answer to the question, "What level of quality is of economic importance to you?" often serves as an excellent guide to sample size. The objective probability chosen, as will become obvious with subsequent discussion, is a function of engineering considerations.

Sample size (attributes), binomial calculation. In the closing paragraphs of Chapter III a brief discussion was given of lot sample size and grand-lot sample size, and two charts were offered (Figs. 3·2 and 3·3) for reading these sample sizes, under a set of definite

[1] Older books on the theory of errors tend to ignore practical distributions; assume the Gaussian or normal distribution as a matter of course, and assume the observed dispersion as the true dispersion. For practical purposes such procedure is likely to be a pleasant delusion. Some recent books on industrial statistics provide the methods for the solution of practical problems on sample size; but, so far as known, none attacks this essentially complicated problem from the viewpoint of getting ready answers to engineering problems.

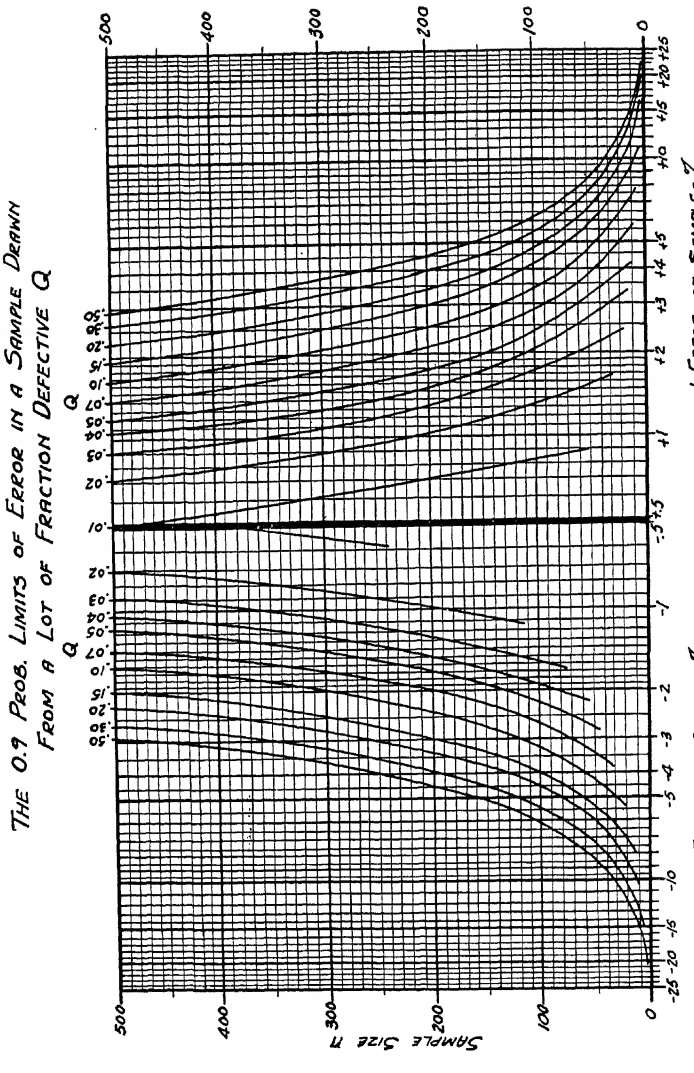

THE 0.9 PROB. LIMITS OF ERROR IN A SAMPLE DRAWN FROM A LOT OF FRACTION DEFECTIVE Q
Q
SAMPLE SIZE n
-ERROR OF SAMPLE-%
+ERROR OF SAMPLE-%

FIG. 10-1.

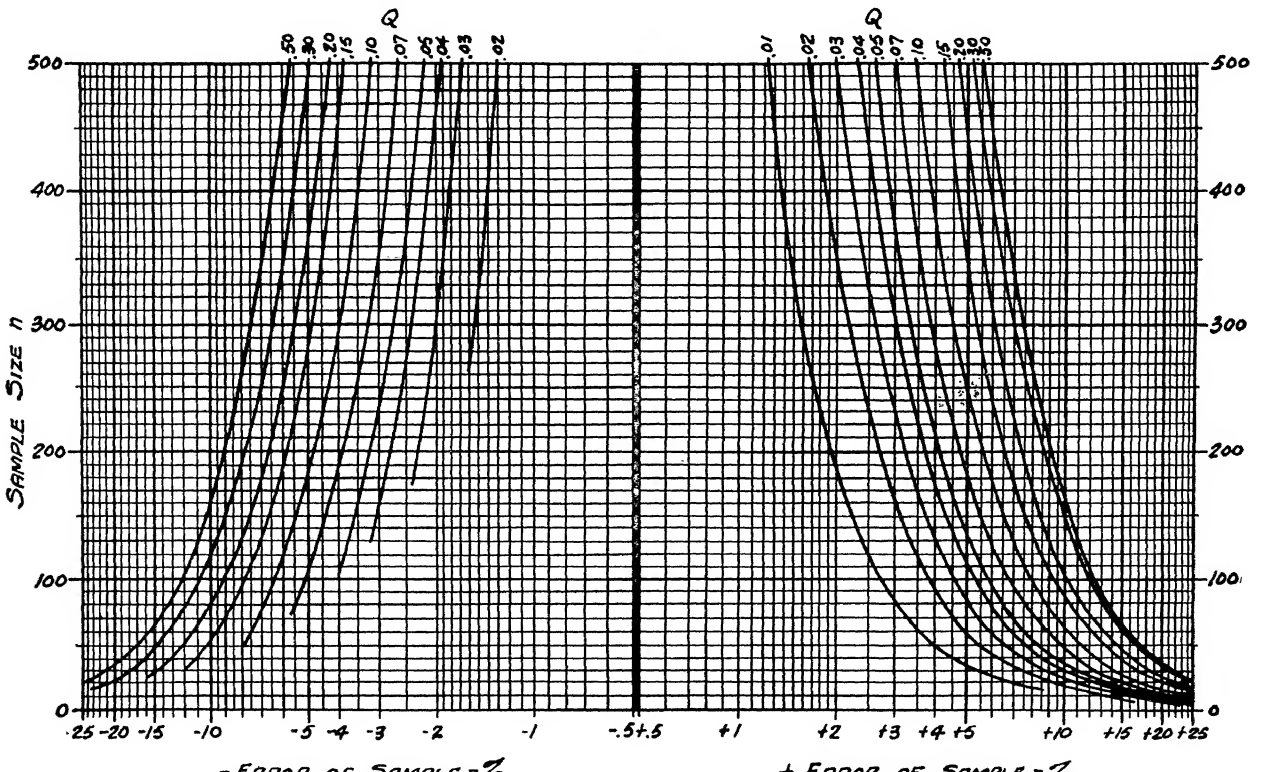

THE 0.995 PROB. LIMITS OF ERROR IN A SAMPLE DRAWN FROM A LOT OF FRACTION DEFECTIVE Q.
Q
.50 .30 .20 .15 .10 .07 .05 .04 .03 .02
Q
.01 .02 .03 .04 .05 .07 .10 .15 .20 .30
SAMPLE SIZE n
500
400
300
200
100
0
-25 -20 -15 -10 -5 -4 -3 -2 -1 -.5 +.5 +1 +2 +3 +4 +5 +10 +15 +20 +25
-ERROR OF SAMPLE - %
+ ERROR OF SAMPLE - %

FIG. 10·2.

criteria suggested in that chapter. Since these criteria are not likely to suit all conditions, answers of a more general nature must be offered.

In general one would like to know what sample size to take in order that the probability will be P that the observed fraction defective in the sample will be correct to within $\pm E$. One can have the probability P, but in general one cannot have the $\pm E$, because the distribution [2] is apt to be skewed; and plus and minus errors of the same magnitude are not equally likely. Therefore, having estimated Q and having chosen P, one can get the answer [3] that, with sample size n, the probability is P that the sample fraction defective will not be greater than $Q + E_1$ and the probability is P that it will not be less than $Q - E_2$.

Figure 10·1 answers this problem for the probability 0.9, various values of Q, and values of n up to 500. For values of Q not shown, one may interpolate. A study of this chart will throw much light upon the behavior of samples by attributes. For example, if the lot Q equals 0.04, and the sample size is 500, the probability is 0.9 that the sample Q will be less than 0.04 + 0.01 and the probability is 0.9 that it will be greater than 0.04 − 0.01; i.e., a sample size of 500 yields a probability of 0.8 that the observed fraction defective will be correct within ± 1%. That is the kind of answer one likes. However, the plus and minus increments are approximately equal only because the sample size is so large as to render the distribution almost symmetric. If one goes on down the $Q = 0.04$ curve, one finds that, for $n = 210$, the limits [4] are + 1.50% and − 1.46%; for $n = 107$, the limits are + 2.00% and − 1.87%; and for $n = 50$, the plus limit is 2.15% and there is no minus limit because the curve for $Q = 0.04$ stopped at $n = 56$ on the negative plot. Why did the curve stop at n equals 56? Because, with $n = 56$ and $Q = 0.04$, the probability is 0.9 for at least 1 defective. If n is less than 56, there cannot be a 0.9 assurance of at least 1 defective. (The probability

[2] The term *distribution* is used to refer to the bar chart frequency diagram; not to $\sigma(x)$, where $\sigma(x)$ is the probability that the random variable is less than x.

[3] It should be noted that this prediction can be made irrespective of conditions of control, requirements as to universe, etc., and is subject to the sole restriction that the samples be random. See the discussion at the beginning of Chapter II and Appendix A.

[4] Of course, one can quote the larger of the two limits with a plus-or-minus sign (in this case, 0.04 ± 0.015), with the interpretation that the probability is 0.8 that the observed sample fraction defective will differ from the lot fraction defective by less than $\pm a$.

of at least 1 defective would have to be slightly less than 0.9). Hence, for lots of $Q = 0.04$, there is no use in bothering with sample sizes of less than 56, if one (*a*) is interested in a probability of 0.9 that the precision is within 2.2% or (*b*) wishes a probability of 0.9 of detecting lots as poor as 0.04 fraction defective, if they exist. This last statement should be compared with the latter part of Chapter III.

Figure 10·2 is a like chart, but for the probability 0.995. It should be noted that the $Q = 0.01$ line does not appear on the negative side. For $Q = 0.01$, the sample size must be 500 to have a 0.995 probability of at least 1 defective. Charts are offered for two probability levels in this type of estimate, just as Charts I_Q are offered for like levels, in order that one may have a low probability where occasional errors are less important than smallness of sample size or when the a priori knowledge is scant (as in the early stages of experimental work), and one may have a high probability when errors are very important or when the a priori knowledge regarding the level of product is very cogent.

These charts were constructed by the process of replotting data from the charts for I_Q. For example, referring to Chart $0.9 = I_Q$ and taking Q equal to 0.02, one finds that, for $c = 1$, $n = 115$; hence with Q equal to 0.02 and n equal to 115, the probability is 0.9 that at least 1 defective will be observed. In this instance, the observed fraction defective is equal to or greater than $1/115 = 0.0087$; hence the probability is 0.9 that the negative error will be less than $0.02 - 0.0087$ or 0.0113. This is the bottom point on the $Q = 0.02$ curve. Successive points are calculated in like manner. Of course, the curves are only approximate, for only whole numbers of defectives can be observed. However, as there is always uncertainty in the estimate of Q, it appears better to draw continuous curves for estimating purposes rather than plot a series of points.

Approximate methods for various probability levels. The charts have been offered to make reasonably precise solutions easy. Ordinarily, the occasional user of statistical methods does not have time for protracted calculations; however, occasion may arise when approximate solutions for other probability levels are desired, even at the expense of some labor.

As shown in Chapter I, all the properties of attributes may be regarded as functions of the binomial $(P + Q)^n$. However, in order to answer the question of what sample size to use when $Q = 0.02$ so that the probability will be 0.8 that the sample fraction defective will

be correct to within approximately ± 0.01, one would have to do a considerable amount of summing of high-powered terms of the expanded binomial. The summing of binomials with large exponents by ordinary methods is hopelessly laborious. Therefore, if n is large, it is common practice to sum the binomial approximately by one of two methods. If Q is small and n reasonably large, one can use the Poisson distribution. If Q is not small (the nearer to 0.5 the better) and n is fairly large, one can use the normal or Gaussian distribution. The latter procedure will be explained under Sample Size (Variables) on page 93 et seq. The former is offered herewith. Both are somewhat tedious.

Sample size (attributes), Poisson calculation. It is shown in elementary textbooks on statistics [5] that, if Q is very small and n very large, the successive terms of the binomial $(P + Q)^n$ can be reduced to

$$e^{-\bar{c}} + \bar{c}e^{-\bar{c}} + \frac{\bar{c}^2}{2!}e^{-\bar{c}} + \frac{\bar{c}^3}{3!}e^{-\bar{c}} \cdots,$$

where the successive terms are the probability of 0, 1, 2, etc., defectives; e is the Napierian base of the logarithm; $i!$ is the product of all the natural numbers from 1 to i inclusive; and $\bar{c}$ is nQ, i.e., the average number of defectives for the sample size taken. Note that the binomial treatment of $(P + Q)^n$, n very large and Q very small, gives rise to the handling of very large numbers. Consider $(0.99 + 0.01)^{200}$. On the other hand, the numbers in the Poisson exponential limit are moderate, since one will not often be interested in a value of $\bar{c}$ that is greater than 4 or 5. Practically all engineers' handbooks tabulate e to various positive and negative powers, and a solution of the above equation would not be difficult. However, Fig. 10·3 shows a chart which should give sufficiently accurate solutions for most practical purposes. This chart gives all probabilities from 0.001 to 0.999. It will be found useful for several types of work.

By way of parallelism with the notation [6] used in Charts I_Q, it is well to view Fig. 10·3 as giving the probability of at least c defectives in a sample (do not mention size), when the average number of defectives in samples of that size is $\bar{c}$. Of course, the sample size should theoretically be infinity, but quite satisfactory results can be obtained in a moderately wide range of finite cases.

[5] See Yule and Kendall, Chapter 10.

[6] Appendix B discusses G. A. Campbell's original solution for Poisson binomial expansion limit over a wide range.

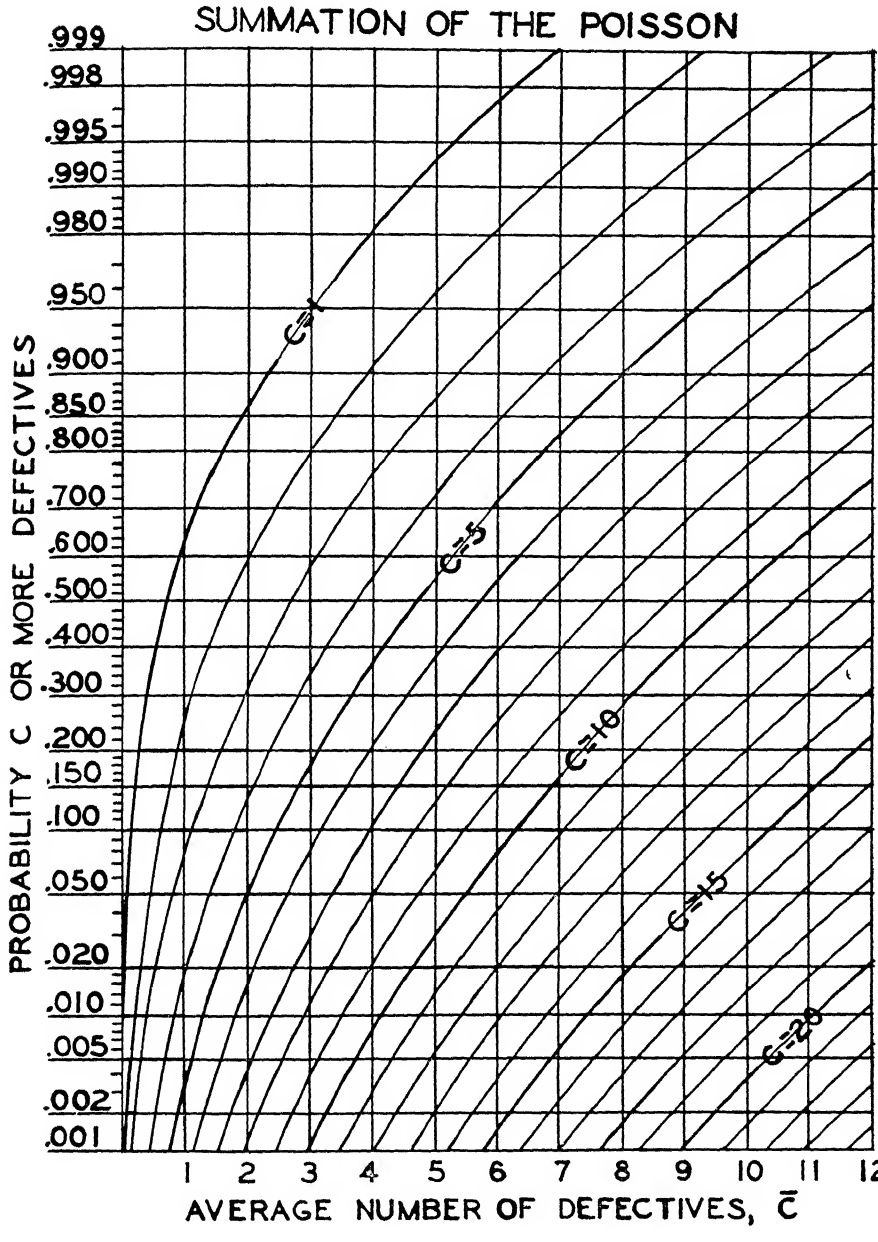

SUMMATION OF THE POISSON
PROBABILITY C OR MORE DEFECTIVES
.999
.998
.995
.990
.980
.950
.900
.850
.800
.700
.600
.500
.400
.300
.200
.150
.100
.050
.020
.010
.005
.002
.001
C=1
C=5
C=10
C=15
C=20
1 2 3 4 5 6 7 8 9 10 11 12
AVERAGE NUMBER OF DEFECTIVES, C̄

FIG. 10·3.

By way of illustration and check of accuracy, consider the problem previously solved by Fig. 10·1. What is the least number of samples that one can take from a lot of fraction defective 0.02, so as to have a probability of 0.9 of observing at least 1 defective, and what is the precision of the result? Entering Fig. 10·3 on the line $P = 0.9$ and finding its intersection with $c = 1$, one would read $\bar{c} = 2.303$, if one could read the chart that closely. Now $\bar{c}$ equals nQ, hence

$$n = \frac{2.303}{0.02} = 115,$$

which does not differ from the answer previously obtained by more accurate procedure. Of course, the probability is 0.9 that the negative error is less than $0.02 - 1/115$ or 0.0113.

One can also readily find the precision of the positive error associated with the specific sample size 115. If the probability is P that one will observe c or more defectives when the average number is $\bar{c}$, it follows at once that the probability is $1 - P$ that one will observe $c - 1$ or less defectives under like conditions; therefore, entering Fig. 10·3 on the line $P = 1 - 0.9$ or 0.1, and finding its intersection with $\bar{c} = 115 \times 0.02$ or 2.30, one finds that this intersection is almost on the line for $c = 5$. One would estimate the value at perhaps 4.8. That is, the probability is 0.1 that 4.8 or more defectives will be observed, because the chart reads that way, viz., at least c. Therefore, the probability is 0.9 that 3.8 or less defectives will be observed under like conditions, viz., sample size 115, $Q = 0.02$. In this instance the observed fraction defective would be $3.8 \div 115$ or 0.033. The positive error is $0.033 - 0.02$ or 0.013. This appears to check very closely with the more exact method. It should be noted that the check could have been made more readily and precisely by using Chart $0.9 = I_Q$ and Chart $0.1 = I_Q$. In order to calculate precision limits as a function of sample size, it would be necessary to resort to some method such as plotting a series of points in the manner of Figs. 10·1 and 10·2.

In the above example Fig. 10·3 gave rather accurate results because Q was small and n large. As one departs further from the ideal conditions of the Poisson, results are less accurate. For instance, consider the upper and lower limits of defectives for $Q = 0.50$, $n = 21$, and $P = 0.9$. Sample size 21 and $P = 0.9$ are selected in order to have a ready check from Charts I_Q.

If $Q = 0.50$, $n = 21$, then $nQ = 11.5 = \bar{c}$. Therefore, entering Fig. 10·3 on the ordinate 11.5 and finding its intersection with $P = 0.900$, one reads $c = 7.7$. The probability, therefore, is approximately 0.9 that 7.7 or more defectives will be observed. This is the lower probability limit that one seeks. The correct value from Chart 0.9 $= I_Q$ is 8.

To find the upper limit, enter Fig. 10·3 on ordinate 11.5 and find its intersection with $P = 0.100$. One reads $c = 16.4$. Therefore, the probability is approximately 0.1 that 16.4 or more defectives will be observed. Note that this is not the upper limit that one seeks. One wishes a limit such that the probability is 0.9 of c or less defectives. However, if the probability is 0.1 of 16.4 or more, then it is $1 - 0.1$ of 15.4 or less defectives. This is the desired upper limit. Its correct value is read from Chart $0.1 = I_Q$ as 13. Thus the Poisson gives 0.9 probability limits of 7.7 and 15.4 as opposed to correct limits of 8 and 13, which is very poor approximation.

For samples of 50 or more and Q's of 0.1 or less, the approximation becomes rather close. For the probability 0.9, $Q = 0.1$, and $n = 50$, the Poisson limits appear to be 2.8 and 7.4 as opposed to Chart I_Q limits of 2.9 and 7.3.

Sample size (variables). Chapter V gave a brief discussion of the distribution of the variable X in a parent universe, the distribution of the averages of samples of size n drawn from the parent universe, and the distribution of standard deviations of samples of size n drawn from the universe.[7] These concepts should be borne in mind when considering sample size. It was further pointed out that the well-known normal distribution is practically never met in practice, and that the distributions of $\bar{X}$ and especially of σ are likely to depart considerably from normality. However, the practice of assuming all distributions to be normal is still too popular and is too important historically for the procedure associated with this assumption to be omitted in a discussion of sample size. Nevertheless, one is gravely warned against quoting exact numerical results derived in this manner, and against attaching great importance to results based on this assumption.

[7] As heretofore, the term *distribution* will be used to refer to the density curve, e.g., the well-known bell-shaped curve in the case of normal law. This terminology is adopted for simplicity and in accordance with popular conception. The reader familiar with mathematical statistics will, of course, understand that reference is made to the first derivative of the distribution, when such derivative exists, and to the bar chart diagram of frequency, in the case of discrete variables. See *Introduction to Mathematical Statistics*, J. V. Uspensky, McGraw-Hill Book Co., New York, 1937.

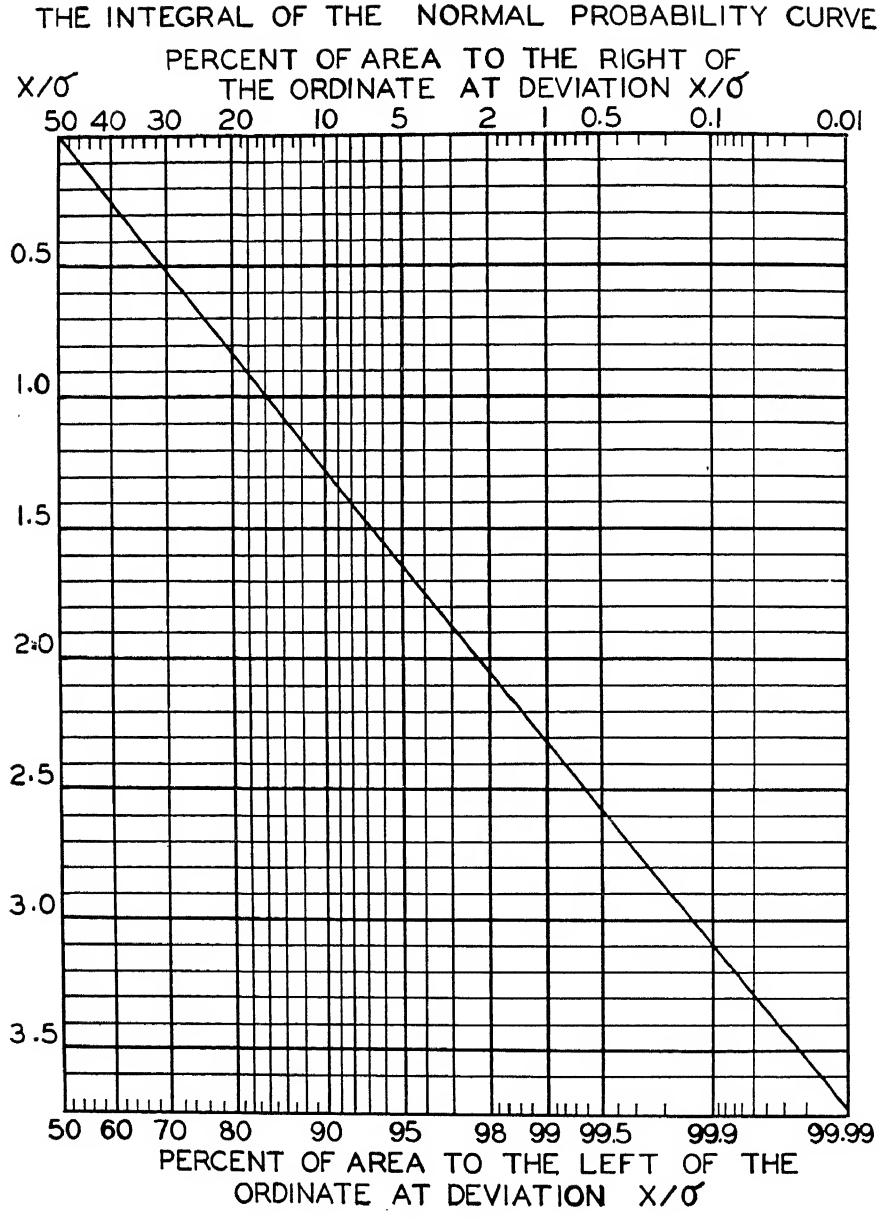

THE INTEGRAL OF THE NORMAL PROBABILITY CURVE
PERCENT OF AREA TO THE RIGHT OF THE ORDINATE AT DEVIATION X/σ
X/σ
50 40 30 20 10 5 2 1 0.5 0.1 0.01
0.5
1.0
1.5
2.0
2.5
3.0
3.5
50 60 70 80 90 95 98 99 99.5 99.9 99.99
PERCENT OF AREA TO THE LEFT OF THE ORDINATE AT DEVIATION X/σ

FIG. 10·4.

Figure 10·4 is offered for reading probabilities under the normal distribution as it is quicker and easier to read than most tables of the normal probability integral. This figure is merely the right half of the bell-shaped curve drawn to such a scale that the curve becomes a straight line. Since the curve is symmetric, the left half is not required. The upper and lower scales show the percentage of area to the right and left of any point on the curve. For example, an ordinate erected at the point $X = 2\sigma$ will (noting the abscissa of the ordinate passing through the intersection of $X/\sigma = 2$ and the curve) cut off an area of 2.3% on the tail of the curve and an area of 97.7% on the body. If $\bar{X}'$ and σ' be used to designate the true average and standard deviation of a normally distributed lot, universe, or population, it follows from the conditions set forth above that one can readily calculate the frequency of individuals of the lot between any pair of limits. This frequency, of course, is the probability that any single individual drawn at random from the lot will be between those same limits.

Sample size for distribution limits under normal law. The probability that a single individual will be between $\bar{X}' \pm t\sigma'$ is readily read from Fig. 10·4. Having read this fraction (probability), all one needs to do is treat it as a fraction defective, and obtain the necessary n for the probability P of at least k defectives, in the manner described for attributes. As a military example, let it be assumed that hits are normally distributed, and that the center of impact is on the center of the target; i.e., aiming is correct and $\bar{X}'$ equals zero. What is the number of trials necessary to have a probability of 0.9 of scoring at least 2 hits, when σ' is 100 feet and the length of target is 135 feet? Only range is considered.

Since the center of impact is on the center of the target, the distance $\pm X$ to its boundaries is $\pm$ 67.5 feet; therefore X/σ equals 0.675. By entering Fig. 10·4 with X/σ as 0.675, one reads that the area to the right of this ordinate is 25%; the area to the left is 75%; and consequently the area between two such symmetrically placed ordinates is 50%. That is, the probability of hitting on a single trial is $\frac{1}{2}$, for the length of target is 2 probable errors.[8] To find the num-

[8] A term which is defined as the error which is as likely to be exceeded as not. It has practically disappeared from statistical usage. In general it is neither an error nor probable. It is very misleading and entirely without meaning unless the distribution is normal. The standard deviation, on the other hand, is always valid. The same can be said of variance (the squared std. dev.), but being non-linear, its interpretation is not as obvious as that of standard deviation.

ber of trials, n, in order to have a probability of 0.9 of 2 or more hits, one merely needs to know what value to assign n in the expression $(0.5 + 0.5)^n$, so that the sum of the terms beyond the second will be equal to 0.9; or, entering Chart $0.9 = I_Q$ with $Q = 0.5$ and finding its intersection with $c = 2$, one reads $n = 6$.

Now let it be noted that this problem did not need to be of a military character. It would be the same problem if one were calculating the percentage of one's manufactured product between certain limits, or the probability that k or more articles in a sample of n would be outside of certain specification or tolerance limits. What is the matter with this solution?

Dangers in the use of the normal integral in practical work. Nothing is the matter with the solution under the conditions stated. The fault lies in the fact that these conditions are seldom found in practice. In the first place, one generally does not know $\overline{X}'$ (in the above case zero) or σ'; but, instead, only $\overline{X}$ and σ, which are estimates of these parameters. This, however, need not be such a serious fault, for if $\overline{X}$ and σ are calculated from even a moderately large number of observations, they are not likely to be greatly in error, and one need only modify his statement by saying that the probability is approximately 0.9 that at least 2 hits will be scored in 6 trials. There is a much more serious hazard in this procedure, however, than that of using estimates of the parameters instead of the parameters themselves.

The use of the integral of the normal probability curve involved the assumption that the parent distribution was normal. The solution was not predicated on the actual distribution. The determination of the actual distribution, in the practical case, with even moderate accuracy would require a very large number of observations. One would likely need something of the order of 1000 to 5000 observations, to, let us say, get a measure of the average, standard deviation, skewness, flatness; and then apply a goodness-of-fit test such as the Chi-square Test.[9] By way of further discouragement, it might be remarked that the distribution, even if determined, would be likely to be too cumbersome to be of ready practical use; and, in general,

[9] It is recommended that the occasional user of statistical methods seek the advice of a competent statistician when the application of more complicated statistical methods appears necessary or desirable. Even such well-known tests as the Chi-square test, the "t" test, and the "z" test may readily be misinterpreted, unless one is thoroughly cognizant of the exact meaning of the test. Furthermore, even the literature specifically devoted to these tests does not necessarily warn one of their limitations.

the occasional user of statistical methods may as well spare himself the trouble. If then, probability predictions predicated on normal law are invalid for the practical problem of measuring the limits of the parent distribution, and if the empirical determination of the actual distribution is also impractical, what shall one do?

No nice answer is going to be found to this question. However, a procedure can be pointed out by which the person applying the statistical methods can generally make valid predictions of a satisfactory form for practical use. In discussing the procedure, it appears best to divide the problem into two parts: (*a*) broad limits of high probability, i.e., X/σ greater than the order of unity; and (*b*) narrow limits of low probability, i.e., X/σ less than the order of unity.

Sample size for distribution limits, X/σ large. If one considers distributions which only roughly approximate the bell-shaped curve of normal law, one sees that these distributions can be considerably skewed, flattened, or peaked without greatly changing the area under the curve which is contained between limits set at plus and minus two, three, or more standard deviations. In fact, the Camp-Meidell inequality, which was cited in Chapter V, gave the following assurance.

$$P_{t\sigma} \geq 1 - \frac{1}{2.25t^2},$$

where $P_{t\sigma}$ is the probability within the interval $\pm t\sigma$. This inequality is subject to the restriction that the distribution be smooth, unimodal (only one hump), with average value coinciding (in practice we may say approximately coinciding) with the most frequent value. If engineering judgment is called into play, one can generally judge from only a moderate number of samples, or even from the kind of article itself, whether or not these conditions are fulfilled. If they are, then one can abandon the precarious ground of normal law, and make valid predictions based on the Camp-Meidell inequality.

Returning to the military example, what is the number of trials necessary to have a probability of 0.9 or greater of scoring at least 2 hits, when σ is 100 feet and the length of target is 500 feet? In this instance, it is known that the distribution does not fully comply with the requirements of normal law, but that it does, at least, satisfy the Camp-Meidell [10] inequality.

[10] "A New Generalization of Tchebycheff's Statistical Inequality," B. H. Camp, *Bulletin of the American Mathematical Society*, Vol. 28, 1922. M. B. Meidell "Sur un problème du calcul des probabilités et statistiques mathématiques," *Comtes Rendus*, Vol. 175, 1922.

With the center of impact on the target, X/σ equals 2.5. Therefore,

$$P_{2.5\sigma} \geq 1 - \frac{1}{2.25(2.5)^2} \quad \text{or} \quad 0.93.$$

Therefore 93% or more of the parent distribution will fall between $X \pm 2.5\sigma$, and the probability that a single trial will be between these limits is at least 0.93. Now, to find the n so as to have a probability of 0.9 of at least 2 hits, one may call the hits defectives, and use charts of I_Q for this purpose. One therefore wishes to find n for $0.9 = I_{0.93}\ (c,\ n - c + 1)$, where $c = 2$. Except for a small range placed thereon for convenience, Chart $0.9 = I_Q$ does not go beyond $Q = 0.50$; hence it will not yield a solution. However, it was not necessary to run the charts beyond the $Q = 0.50$ because the following relationship exists:

$$I_Q(c,\ n - c + 1) = 1 - I_{1-Q}(n - c + 1,\ c).$$

Therefore, one can consult Chart $0.1 = I_Q$ for n in the case of $I_{0.07}$ $(n - c + 1,\ c)$ and find n for $c = 2$, observing that the terms in the parenthesis representing n and c have been interchanged. Therefore, entering Chart $0.1 = I_Q$ on ordinate 2, and finding its intersection with $Q = 0.07$, one reads 0.15 as a c line. The note at the top of the chart says that c's are shown as 1 less than their real values; hence this value is 1.15. Therefore, the first term in the parenthesis of $I_{0.07}(n - c + 1)$ is 1.15, or $n - c + 1 = 1.15$. Therefore, $n - 2 + 1 = 1.15$. Therefore, $n = 2.15$. Hence, one has a choice of 2 trials with a probability of perhaps slightly less than 0.9, or 3 trials with a probability [11] of somewhat greater than 0.9 It should be noted that the single trial probability (see Fig. 10·4) under normal law would have been 0.988 and that consequently 2 trials would have appeared to have more than satisfied the requirement of a 0.90 assurance, since the probability of exactly 2 hits out of 2 trials is $(0.988)^2$ or 0.976.

If one cannot be assured that the requirements of the Camp-Meidell inequality are met, one can fall back on the Tchebycheff

[11] The reader should not get discouraged if the manipulation of the incomplete beta-function ratio appears to present difficulties. Unless used constantly this process is almost certain to be provokingly tedious. The practical solution lies in the construction of special-purpose charts to fit one's special need, when use is frequent.

inequality. It holds irrespective of the form of the universe and tells one

$$P_{t\sigma} \geq 1 - \frac{1}{t^2}.$$

Applying this inequality in the above example, one has

$$P_{2.50\sigma} \geq 1 - \frac{1}{(2.5)^2}, \quad \text{or} \quad 0.84.$$

Having obtained the frequency within the specified limits, one would solve for n as previously outlined.

Limits within which practically all of a distribution will lie. An important inversion of the question of sample size consists of estimating the limits within which practically all of a lot or universe of effects will lie. This question frequently arises in connection with specifications, since it is not unusual for a specification to require that all of a product shall lie between certain limits.[12] The question also arises in problems of design, when it is necessary to make allowance for maximum and minimum products. From the discussion of Chapter IV, the remarks of this chapter, and Table 10·1, it is obvious that practically all of a product can be expected to lie within plus or minus about three or four standard deviations, depending upon how closely the distribution appears to approximate normality. It is also obvious that it is advisable not to quote any numerical percentage but merely to say "practically all."

It should be noted that the adoption of the probabilities furnished by the inequalities, when the ratio of X to σ is large, does not drastically reduce the figure which would be obtained from the integral of the normal curve and has the important advantage of leading one to results of which one is sure, instead of dubious or false results for which one may have to render subsequent apologies. The following tabulation gives basis for comparison.

[12] The requirement is, of course, absurd where inspection is destructive and operational verification of the requirement is neither contemplated nor possible. Such a specification is therefore without definite meaning either to purchaser or vendor. Hence, it is obvious that in the interests of clarity and fairness the acceptance specification should be confined to a simple statement of (1) the quantity and (2) the kind of evidence which will be accepted as a satisfactory indication that the product will meet the chosen quality standard. Such a statement must include method of sampling, sample size, and limits of parameters. The subsequent discussion will indicate a way in which (1) and (2) may be stated so as to yield a chosen probability that practically all of the product will be between certain limits.

TABLE 10·1

PROBABILITY FOR X/σ, YIELDED BY DIFFERENT TECHNIQUES

Technique \ X/σ	0.667	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Normal law...	0.4845	0.6827	0.8664	0.9545	0.9876	0.9973	0.9994	0.9999
Camp-Meidell.	0	0.5556	0.8025	0.8889	0.9338	0.9520	0.9638	0.9723
Tchebycheff..	0	0	0.5556	0.7500	0.8400	0.8889	0.9185	0.9375

Sample size for distribution limits, X/σ small. It is obvious from Table 10·1 that the Camp-Meidell inequality gives no assurance of any frequency between limits of less than approximately 1 probable error and the Tchebycheff inequality fails at plus or minus one standard deviation.[13] In this dilemma shall one turn to time-honored normal law for lack of something better?

"All right," said the affable guest to the host's small daughter, "I will tell you a story about Africa. Have you ever been to Africa?"

"No, sir," said the child.

"Ah! I may speak freely then."

Probability predictions of doubtful validity have been made in many fields, under the favorable conditions which surrounded the guest, and have brought no penalty because they were of such a nature that they were not subject to future verification. In the industrial field thousands of inspections of a product can be expected to be performed in the next few months or the next year, and the industrial statistician who makes a practice of accepting the hazard of making invalid predictions regarding the subject matter of those inspections is certainly (not probably) going to get caught.

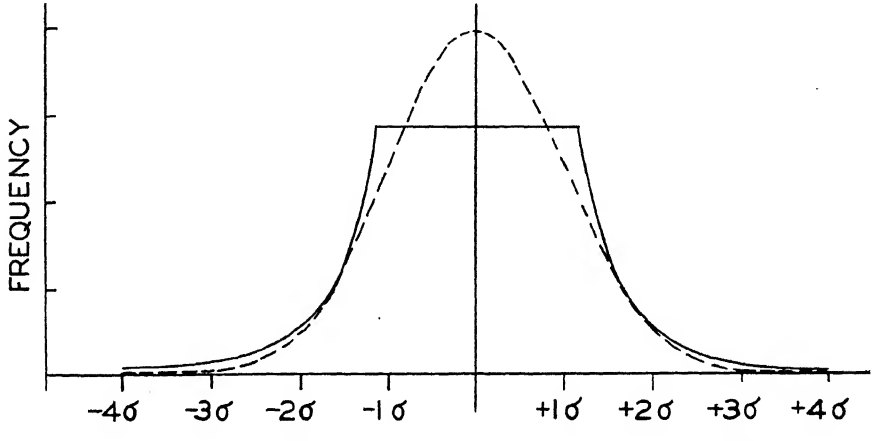
There is generally a way to make valid predictions of frequency within limits even with X/σ small. In this case, one need not know the shape of the whole distribution but only that within approximately plus and minus one standard deviation. This type of knowledge does not require huge samples. Knowledge of the whole distribution required a huge sample, because frequencies in cells near the tails of the distribution are very small, and a huge sample is required to yield a sensible frequency in these regions. The frequency for cells near the

[13] This observation should kill the last vestige of faith in the once widespread but now rapidly passing uses of probable error.

middle of the distribution is large, and that portion of be empirically approximated with much less effort.

However, even such approximation may not be absolutely necessary. If the quality characteristic under investigation has been subject to any material amount of observation, one almost always has knowledge of the type that frequencies at distances of less than $t\sigma$ from the mean are never less than frequencies at $t\sigma$ (t being small).

Again, returning to the military example of scoring hits on the target, in which instance σ was 100 feet, one may know that the frequency of hits on intervals that are referred to the center of impact



SCHEMATIC DIAGRAM OF A PARTIALLY KNOWN DISTRIBUTION

FIG. 10·5.

by $\pm$ 0 to $\pm$ 10 feet, $\pm$ 10 to $\pm$ 20 feet, $\cdots$ $\pm$ 80 to $\pm$ 90 feet is never less than the frequency of hits on the interval at the distance $\pm$ 90 to $\pm$ 100 feet (like signs taken together). In particular, this condition holds rather closely for every method of projecting projectiles at targets which is in ordinary use. Therefore, in the absence of even enough data to construct the central portion of the distribution curve, one can be assured of at least a chosen uniform density of probability within $\bar{X} \pm t\sigma$. The distribution may therefore be regarded as shown schematically in Fig. 10·5. The solid curve resulting from the application of the Camp-Meidell inequality must be regarded as a mean between the frequencies summed at symmetric intervals, because the inequality guarantees nothing on either side of

the mean, but only the sum between symmetric limits. The dotted curve, of course, shows the normal curve, for purposes of comparison. It should be noted that, if the actual distribution is known to be somewhat irregular, and also if its irregularities cannot be controlled

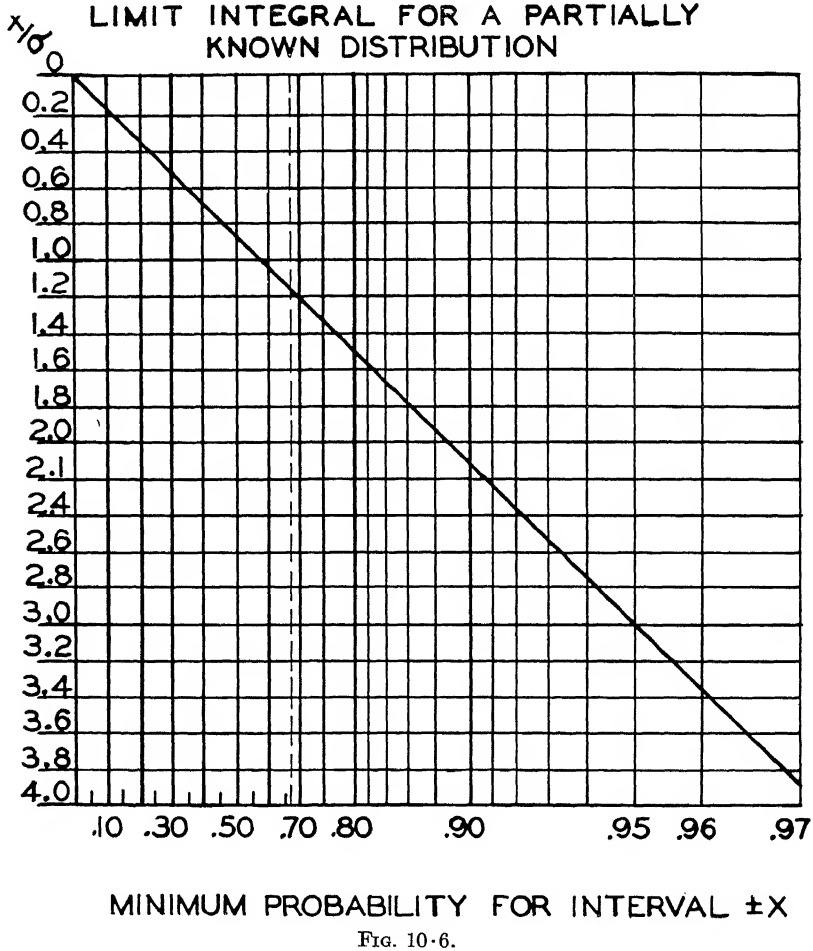


FIG. 10·6.

with respect to practical applications, the actual distribution, if obtained, is not likely to yield probability assurances significantly higher than those given by this rather simple analytical scheme. Figure 10·6 gives the integral of this minimum probability curve. Of course, similar construction is easy for different limits of uniform probability and for the Tchebycheff inequality.

Sample size for limits of the average. Calculations regarding averages involve considerably less incertitude than those regarding the individuals of the population because: (1) averages of samples of size n drawn from a normally distributed population are themselves normally distributed; (2) averages of samples of size n drawn from a non-normal population approach normal distribution as n increases; and (3) the standard deviation of the distribution of averages, $\sigma_{\bar{X}}$, is equal to $\sigma'/\sqrt{n}$. It is an easy problem in mathematical statistics to show that statement (3) holds, irrespective of the form of the parent population, subject to the sole restriction that the parent distribution has a finite first moment, which is a condition that could hardly fail to be met in practice. In calculations regarding averages, one can therefore apply normal law with a moderate degree of safety; [14] however, it is still the part of wisdom not to quote numerical probabilities, but merely to assign zones of probability, e.g., extremely high probability of the order of 95 to 100%; high probability of the order of 90 to 95%, etc. Therefore, if one has a satisfactory estimate of σ', and selects an order of magnitude for the probability, one can readily calculate the sample size required in order that the probability shall be of the order of P that the average of the sample of n shall be within $\pm E$ of the true average. To do this, one need only ascertain the ratio of X/σ associated with P from Fig. 10·4, and solve the equations:

$$E = t\sigma_{\bar{X}},$$

$$\sigma'_{\bar{X}} = \frac{\sigma'}{\sqrt{n}}$$

or

$$n = \left(\frac{t\sigma'}{E}\right)^2.$$

For example, suppose that one is going to test a lot of fuzes to determine the average burning time. Predecessors of this lot have showed control with a standard deviation of the order of 0.03 second, under standard test conditions. What sample size must one take in order to have a high probability (probability of the order of 90 to 95%) that the observed average will be within $\pm$ 0.01 second of the true average?

[14] The restriction that the population be a population, and not a hodgepodge of populations, as assured by prior knowledge or demonstrated by satisfactory response to a test by a control chart, is, of course, not removed.

Figure 10·4 shows that 95% corresponds to 2σ under normal law. Figure 10·6 shows that 90% corresponds to 2σ under Camp-Meidell. Two, therefore, appears to be an appropriate value of t. Substituting, one gets

$$n = \left(\frac{t\sigma'}{E}\right)^2 = \left(\frac{2 \times 0.03}{0.01}\right)^2 = 36.$$

If, on testing the sample, the standard deviation should appear to be significantly greater than 0.03 second, one might then conclude that some additional samples should be tested in order to yield the desired precision of results.

Sample size for limits of the standard deviation. Calculations regarding the standard deviation can be made only on less certain grounds than those regarding averages because: (1) standard deviations of samples of size n from a normally distributed population are not themselves normally distributed, but only approach normality as n increases; (2) the distribution of standard deviations of samples of size n drawn from a non-normal population cannot, in general, be satisfactorily handled on theoretical grounds, although there are empirical grounds for considerable assurance that under controlled conditions such distributions approximate normality as n increases. The application of normal law is therefore quite dubious; and, instead, it is recommended that probabilities be quoted only by categories such as highly probable and fairly probable, and that in these instances the upper limit be taken from the Camp-Meidell inequality and the lower from the Tchebycheff. The relationship between the universe standard deviation, σ', and the standard deviation of the standard deviation, is $\sigma_\sigma = \sigma'/\sqrt{2n}$. Therefore,

$$n = \tfrac{1}{2}\left(\frac{t\sigma'}{E}\right)^2$$

Hence, if one has a satisfactory estimate of σ' and obtains an appropriate t for the order of magnitude of probability he chooses to select, one can confidently calculate the sample size, if the product is controlled.

For example, in testing the same lot of fuzes just cited under averages, suppose that one wishes to know what sample size to take in order that the probability shall be high (probability of the order of 0.89 to 0.95) that the observed standard deviation will be within $\pm$ 0.01 second of the true standard deviation? Note that this is the same assurance and same limits demanded under averages.

Figure 10·6 shows that 95% corresponds to 3σ, under the Camp-Meidell inequality. Substitution in the formula $P_{t\sigma} \geq \left(1 - \frac{1}{t^2}\right)$ shows that 89% corresponds to 3σ. Therefore, $n = \frac{1}{2}\left(\frac{t\sigma'}{E}\right)^2 = \frac{1}{2}\left(\frac{3 \times 0.03}{0.01}\right)^2 = 40.5$. The sample sizes for like assurance of like precisions in average and standard deviation are therefore not greatly different. It should be noted that meticulous precision in calculations of this sort is not generally warranted, as the actual selection of sample size is often strongly influenced by practical considerations such as convenience, relationship to other tests, number that are ordinarily packed in a box, etc. That is, having determined the theoretical sample size, one will generally select the next larger sample size which is a multiple of standard packing; or, if a multiple is just slightly less than the theoretical sample size, it may sometimes be selected. However, the theoretical sample size is none the less very important as a guide to action.

Sample size (attributes), normal probability calculation. It was stated in the discussion of approximate methods of calculating sample size—sampling by attributes—that an approximate method of calculation of sample size for various degrees of probability in the case of Q of the order of 0.5 would be given under variables. In view of the previous discussion, the method is almost obvious. If Q is approximately 0.5 and n moderately large, the binomial $(P + Q)^n$ can be closely approximated by normal law. In this case $\overline{X} = nQ$ and $\sigma = (PQn)^{1/2}$, or, in terms of fractions, $\overline{X} = Q, \sigma = (PQ/n)^{1/2}$. Since normality is assumed, plus and minus deviations of like magnitude must be assumed to be equally probable. Thus, having an estimate of Q, one has like estimates of $\overline{X}$ and σ, and consequently all types of calculations can be made under normal law.

For example, if the fraction defective is 0.50, what sample size must one take in order to have a probability of 0.8 that the observed fraction defective is not in error by more than ± 5%?

Here one wishes to find the ratio of $X/\sigma = t$, which in normal law corresponds to 90%; i.e., the probability is 0.9 that the fraction defective will be less than $\overline{X} + t\sigma$, the probability is 0.9 that the fraction defective will be greater than $\overline{X} - t\sigma$, and the probability is 0.8 that the fraction defective will be between $\overline{X} \pm t\sigma$. Figure 10·4 shows this value of t to be 1.28.

Therefore,

$$0.05 = 1.28 \sqrt{\frac{(0.5)(0.5)}{n}}$$

and therefore

$$n = 164.$$

Figure 10·1 shows this answer to be correct. In like manner, the c on Chart 0.1 $= I_Q$ for $n = 164$, $Q = 0.50$, is 90, and $90/165 = 0.549$, which is just within 5% of 50%. However, if this same procedure is applied when Q is very small, and solved under calculation by the Poisson, results will not be so good. There, Q was equal to 0.02, and the 0.8 probable limits of error were + 0.013 and − 0.0113. Solution by the method of normal law for sample size 115 gives

$$E = \pm 1.28 \sqrt{\frac{(0.02)(0.98)}{115}} = \pm 0.017,$$

which is a poor approximation of 0.013. If n had been smaller, the approximation would have been poorer.

Thus, with Q of the order of 0.5, approximate calculations for any probability can be made by means of normal law; with Q small, the same is possible with the Poisson; but n should be large in both cases. For intermediate values of Q and n small, there is no approximate method, and one must use the binomial. The most satisfactory method appears to consist of all solutions from the binomial summation as given by charts of the incomplete beta-function ratio like Figs. 10·1 and 10·2, which can be prepared for any probability.

Unit sample size in the special case of indeterminate sample size. Chapter VIII discussed a special type of sampling where inspection is made for defects per article and the presence of defects does not necessarily result in the classification of the article as a defective article. Attention was called to the applicability of this type of inspection to surfaces such as photographic film, textiles, plated products, etc. The sample size was indeterminate, for the reason that a single unit area of surface may have a potential capacity of many defects, although zero or only a few may be present. During the discussion a significant fact of sample size was brought out. If the unit area selected as a sample proved to have such a small average number of defects, $\bar{c}$, that there was not a reasonably good probability of observing at least 1 defect per sample of marginal quality, then the unit area should be increased.

In the present chapter interest centers in this type of sampling only to the extent of answering the question, "How many unit samples must I take in order that the probability will be P that the observed average number of defects per unit sample, $\bar{c}$, does not differ from the true average number of defects per unit sample, $\bar{c}'$, by more than E?" It was noted in Chapter VII that the universe of defects has the Poisson distribution. In the Poisson, the mean is $\bar{c}$ and the standard deviation is $\bar{c}^{1/2}$. The standard deviation of the average is, of course, the square root of the average number of defects divided by the square root of n, since this relationship holds if the distribution has a finite first moment. Also, the distribution of averages must be approximately normal and must approach normality rapidly as n (n being the number of unit samples) increases. Therefore, $\bar{c}$ can be regarded as normally distributed with mean equal to $\bar{c}$ and standard deviation equal to $(\bar{c}/n)^{1/2}$. It follows that n can be calculated, in this case, in exactly the same manner as sample size for limits of the average.

$$n = \left(\frac{t\sigma'}{E}\right)^2 = \left(\frac{t\bar{c}'^{1/2}}{E}\right)^2.$$

For example, suppose that one wishes to know how many sample units must be inspected in order that the probability will be high (probability of the order of 90 to 95%) that the observed $\bar{c}$ will not differ from $\bar{c}'$ by more than $\frac{1}{4}$ of a defect. Suppose that it is known from previous work that $\bar{c}'$ is of the order of 4. The value of t corresponding to $P = 0.95$ is 2 for normal law, and the value of t corresponding to $P = 0.90$ is 2 for the Camp-Meidell inequality. Therefore,

$$n = \left(\frac{2 \times 4^{1/2}}{\frac{1}{4}}\right)^2 = 256.$$

If it is desired to express E as a fraction of $\bar{c}$, then σ' should also be divided by $\bar{c}'$, and the formula becomes

$$n = \left(\frac{t}{E\bar{c}^{1/2}}\right)^2$$

For example, changing E in the above case from $\frac{1}{4}$ of a defect to 6.25%,

$$n = \left(\frac{2}{0.0625(4)^{1/2}}\right)^2 = 256.$$

The most economical sample size. The preceding paragraphs have discussed the sample size required in order to have a given probability that the error in an estimate based on a sample shall be within certain limits. Although a valuable guide to judgment, such information is not of necessity of direct assistance in selecting the sample size which will result in the greatest monetary gain.

To determine the most economical sample size, one gets an expression for all the costs, C, associated with the inspection and the consequences thereof, including the cost of testing and the average cost of misgrading as a function of sample size, n; takes the derivative of C with respect to n; equates the derivative to zero; and solves for n. From a practical point of view there are two difficulties in such a procedure: (1) it is difficult to estimate accurately the costs resulting from the misgrading of a lot; (2) without the aid of suitable charts or tables, which are not at present generally available, it is tedious to evaluate the derivative of C with respect to n, as a function of n. This is not an elementary procedure,[15] and the occasional user of statistical methods had best employ the services of a competent mathematician. However, the reader should at least be cognizant of the existence and general mode of operation of such techniques; and therefore some results of procedures will be briefly outlined as a guide for the occasional user of statistical methods.

For example, suppose that lots of articles are being inspected by attributes; let

n = sample size.
c = number of defective articles in a sample.
N = number of articles per lot.
T = cost of selection and inspection of an article.
M = cost of misgrading an article.
P_m = probability of misgrading a lot of articles.
b = an action limit such that the lot will be rejected if in an n round sample $\frac{c+1}{n+2} \geq b$.
C = cost of inspection plus average cost of misgrading.

Then

$$C = nT + P_m NM,$$

[15] These solutions were proposed by Mr. R. H. Kent, who, in an unpublished report, supplied various charts for the ready selection of the most economic sample size for conditions of tests and costs associated with an ammunition problem.

since the average cost of misgrading is the product of the probability of misgrading, P_m, into the cost of misgrading a lot, MN. To find the value of n for which C is a minimum,[16] take the first derivative of C with respect to n and equate to zero.

One then gets P_m as a function of n, and thus it is possible to evaluate [17] the most economical sample size. Like solutions are possible for sampling of variables. Such types of solution are sometimes helpful when at least approximate limits of sample size are not indicated by some one or more obvious common-sense factors.[18] But in the majority of cases some one or two considerations (such as a limit which must be met with a probability P) are of such predominant importance that a little engineering judgment, in connection with the simple aids previously given, will show that the economic sample size must lie between limits which are so close that refinement by nice formulas is not required.

Reduction of sample size by tests of increased severity. It may be observed that the necessary sample size for accomplishing one's objective is sometimes quite large. This is especially true in sampling by attributes when Q is very small. There is a rather obvious reason for this. If the frequency of an event, e.g., failure, is very small, say 2 or 3 failures in 100, then a large number of trials must be made to determine its frequency with even a moderate degree of precision, because the chance occurrence of 1 or 2 events more or less than average will play such havoc with the observed ratio. It follows

[16] $\dfrac{dC}{dn} = T + \dfrac{dP_m}{dn} NM = 0$, and $\dfrac{dP_m}{dn} = \dfrac{-T}{NM}$.

[17] The evaluation of dP_m/dn may be difficult. If the fraction defective is small, it can be assumed that the Poisson applies, and dP_m/dn reduces to

$$; \left[\frac{b\bar{c}^{c-\frac{1}{2}}}{(c-\frac{1}{2})!} - \frac{Q\bar{c}^{c-1}}{(c-1)!} \right],$$

where $b = (c+1)/(n+2)$. If Q is large, normal law may be assumed and dP_m/dn reduces to

$$\frac{e^{-y^2/2}}{(2\pi)^{\frac{1}{2}}} \left[\frac{y}{2n} - \frac{Q-b}{(PQn)^{\frac{1}{2}}} \right],$$

where $y = (\bar{c} - c + \frac{1}{2})/(PQn)^{\frac{1}{2}}$.

In order to get solutions for n, curves of constant Q should be plotted on a plane of which the coordinates are MN/T and n, where $MN/T = -1/dP_m/dn$.

[18] In *Control of Quality of Manufactured Product*, Chapter XXIII, Shewhart gives an interesting deduction on minimizing the cost of measurement. In this case he distinguishes between the number of articles sampled, n_1, and the number of measurements on each article, n_2. He finds the value of n_1 and n_2 which render the cost of inspection a minimum.

therefore that for economy of sample one should test in a range which yields a sensible number of failures, i.e., shift the test point to a place where the proportion of failures becomes large enough that a difference from some expected proportion can be more readily sensed, care being taken, of course, to see that a shift in test point does not involve a change in engineering features that would vitiate the test.

Engineers have long observed this principle, but on rather different grounds, and not to the full extent. For example, given a certain piece of apparatus which in a working system composed of several pieces of apparatus must not fail at some certain maximum impact, the engineer would generally test such apparatus not at that maximum impact, but at some greater impact. His reason for testing at the greater impact, however, would be predicated on the need of a factor of safety, not on the statistical necessity of observing a considerable number of failures in order to distinguish between levels of quality. Now suppose that practically all articles tested pass this test of increased severity. One would say then that the article appeared to be fit for service, because all (or at least the samples tested) passed the service test plus a factor of safety. However, one would not know whether manufacturer A's product was better than B's or C's; and, if the article happened to be one which was subject to deterioration in service, failures in the poorer quality of product might subsequently become a matter of significance. This holds even in the case of 100% testing. If the test happens to be destructive, thereby necessitating percentage inspection, another difficulty arises. The condition of practically no failures precludes assurance of control or detecting lack of control. Therefore, a quite non-uniform product may escape detection; and the fact that the sample passed the test gives poor assurance that the untested remainder may not contain pieces of apparatus which will not even meet the service condition. It should be remembered that, without a state of control, prediction from sample to lot is invalid.

If, on these grounds, it is accepted that a test of increased severity is sometimes desirable, the question arises as to how severe the test should be. It appears that the test would be statistically most advantageous if made at the point which would yield approximately 50% failures in a product of standard quality; but this observation, even if true, is not sufficiently inclusive, for a knowledge is also needed of what happens at 40% failures, 30%, etc. A brief analytical inquiry therefore seems appropriate.

In the practical application of statistical methods one is interested in learning about quality characteristics of lots of articles. Quality characteristics are sampled by attributes, not because they are per se of a two-category nature, but (*a*) because we choose to classify them in two categories, or (*b*) because our method of measurement is not sufficiently sensitive to classify them in greater detail, or (*c*) because an application of the measure changes the article so as to preclude subsequent measures which would lead to a quantitative measure of its value rather than a mere qualitative measure. In fact, it is almost safe to assume that a quality characteristic in almost any physical population is distributed on a continuous scale as indicated schematically in Fig. 10·7.

SCHEMATIC DIAGRAM OF THE DISTRIBUTION OF A QUALITY CHARACTERISTIC

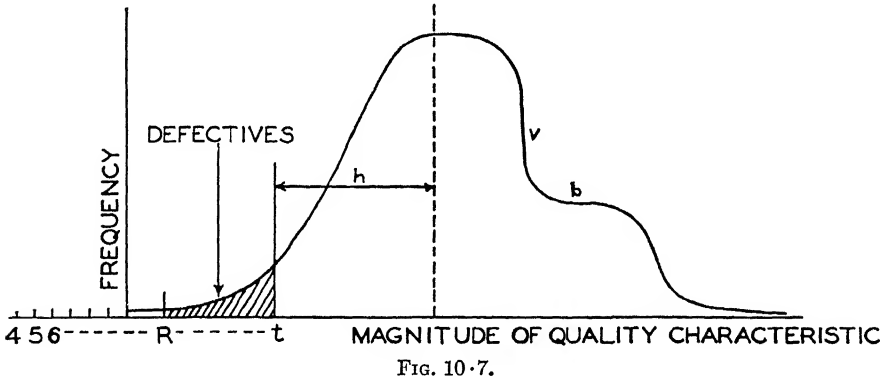


FIG. 10·7.

This may be pictured as the distribution of the quality characteristic resistance to impact, where R is the required amount of resistance to impact and t is the point of test. The distance $t - R$ represents the factor of safety. Any article tested which possesses the quality characteristic to the extent j ($j \geq R$) merely passes the test. The additional extent of its ability to withstand impact remains unknown. The technical objective of a test of this character consists in an attempt to estimate the percentile of the distribution cut off by the point of test, where percentile means the percentage of area to the left of the point of test. In the terminology of attributes, this is the fraction defective at the point of test. In the present case R, of course, should be to the left of the zero percentile. The test point t (including the factor of safety $t - R$)

would, in general, be at some very small percentile, for it is likely that, under customary procedure, only a small fraction defective would be admissible even with the factor of safety. The question for investigation is the relative advantage of moving t to some other point, $t + h = i$, when i is the ith percentile.

An answer to this question with respect to distributions in general is impossible. However, an answer to the question with respect to the normal curve will throw much light on what the answer would be with respect to smooth unimodal distributions which approximate the normal curve.

The solution with respect to the normal curve is easy. The advantage of testing at any percentile is proportional to the precision constant at that percentile or inversely proportional to the standard deviation of the percentile. The standard deviation of the percentile is well known.[19] However, for engineering purposes, it appears better to make the precision a constant and show the relative sample size required at various percentiles to render the same precision. Thus, taking the 0.5 percentile as a base, Fig. 10·8 shows the sample size required for like precision at any other percentile. For example, if a certain degree of accuracy is attained with sample size n at the 0.5 percentile, it will take a sample size of approximately 1.7 times n to attain like accuracy if testing is done at the 0.9 percentile, and approximately 64 times the sample size if testing is done at the 0.999 percentile. It is also seen from Fig. 10·8 that efficiency of test as a function of point of test remains almost constant from the 0.25 to the 0.75 percentile. Thus, considerable latitude in choice of working range is allowable without serious impairment of efficiency.

The advantages of testing in a range which exhibits a sensible number of failures have been covered in detail, because it is a principle which frequently passes unrecognized. However, the principle also involves some dangers, and great care must be exercised in its application. For example, suppose that a distribution happened to have a long range of uniform density of probability as shown at b on Fig. 10·7, or a range rapidly changing density as shown at v, and suppose that one of these ranges was in the vicinity of the testing point. Obviously some invalid conclusions might result. It is therefore obvious that, in applying a test of increased severity, two things should be done: First, the fraction defective of articles of standard quality at the test point should be experimentally established, and,

[19] Chapter XVII, eleventh edition, Yule and Kendall.

at the same time, some experimental verification should be obtained to support the inference that lots of larger fraction defective at the test point are in fact inferior. Second, before attempting to grade articles in a quantitative way, at least the approximate relationship between fraction defective at the test point and the distance between test point and the point of practically no failures should be worked out experimentally for at least a few values of the average.

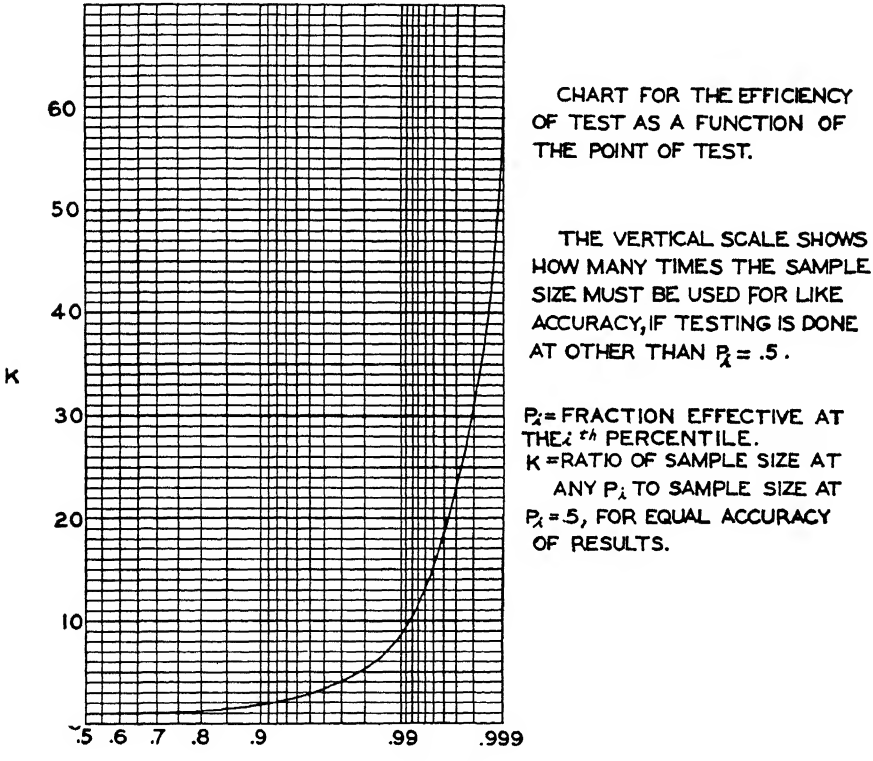


FIG. 10·8.

It is obvious that the application of a test of this character implies that the distribution retains essentially the same shape from lot to lot, that the standard deviation of the distribution does not change significantly from lot to lot, and that the change, if any, is likely to be attributable to a change in average value. This is a broad field, but the control chart technique previously described is sufficiently powerful to leave little to chance if it is employed in connection with tests of this kind.

There are times when a test of increased severity is practically a necessity. For example, the small primer which goes in rounds of ammunition must almost never fail. If sufficient samples from each lot of primers were tested at the degree of initiation expected in the gun to obtain a satisfactory assurance of practically no failures, there would simply be no primers. All lots would be exhausted in test. However, for a certain type of primer it can be determined that lots which exhibit only 1 or 2 failures in 1000 (this from history) show $100P\%$ at a certain test level of initiation (this from trial of the remnants of the historically satisfactory lots). In like manner, data are gathered with respect to various lots that have a less satisfactory technical history, to show that a shift in mean gives expected results. Thus a standard test procedure of increased severity can be established. This type of procedure has proved very satisfactory, both as a test criterion for service and as a method of spotting changes in level of quality. This procedure has also proved effective in process inspection, where destructive tests of piece parts are involved. Each instance, however, appears to require good engineering judgment, statistical study, and a careful experimental check. Statistical methods cannot give something for nothing. They are mere auxiliary tools for facilitating the accomplishment of tasks which could be accomplished at somewhat greater effort without them.

CHAPTER XI

SIGNIFICANT DIFFERENCES

> Appearances to the mind are of four kinds. Things either are what they appear to be; or they neither are, nor appear to be; or they are, and do not appear to be; or they are not, and yet appear to be. Rightly to aim in all these cases is the wise man's task.
>
> —EPICTETUS: *Discourses.*

General discussion. Significant differences are associated with the problem of deciding whether one shall attribute a disparity in sampling results to chance or to an actual difference in the universes under consideration.[1]

This is a problem of great frequency. One is constantly faced with the question, is product type A better than product type B? The basic evidence for solving the problem almost always consists of samples. As a matter of economy, and often of necessity, the samples may be quite few. For example, in development work one may test 15 or 20 tool-made samples of a product. Suppose that 3 fail, and that this is considered unsatisfactory. A change is made in design; another sample of like size tested; and none fail. Questions: Is one design superior to the other? Or may the disparity in sampling results be attributed to sampling fluctuations? Engineers are all too prone to accept evidence of this character as conclusive evidence of a real difference. If one goes back to the principles of Chapter I, it is readily seen that, if a product is 10% defective, the probability of 0 defectives in a sample of 20 is 0.12, whereas the probability of 3 defectives is 0.19. On these grounds, it appears not particularly unlikely that the 2 samples might have come from products of practically the same, if not the same, quality. The engineer is not often brought to account for this sort of mistake in judgment, for he is

[1] If there is a considerable number of samples, and one wishes to discover if 1 or more of them appears to be significantly different from the rest, the control chart methods of Chapters III, IV, VI, VII, and VIII may be applied. The present chapter is devoted primarily to significant differences in 2 small samples of size n, or of size n_1 and n_2, although the principles given can, of course, be applied to several samples if considered successively, or in combinations.

like the doctor who is so frequently accused of burying his mistakes. The engineer's abandoned designs do not arise to mock him. The loss, however, may be nonetheless real. In more general form, the problem can be illustrated thus: having observed c_1 defectives in a sample of n_1 articles, and having observed c_2 defectives in another sample of n_2 articles, shall one's action be predicated on the hypothesis (*a*) that the disparity in sampling results is due to chance, the parent universes being the same, or (*b*) on the hypothesis that the parent universes are really different? Of course, like problems arise with respect to the observed disparity in averages, standard deviations, and other statistics. In order to arrive at a decision, one would like to know the probability that the universes are the same. Now let us consider three important factors in such a decision: (*a*) the level of significance, or degree of objective probability, (*b*) the nature of the probability, and (*c*) the statistical method.

In significant differences, a level of significance is a degree of objective probability. However, the subject matter of the probability is not (as is often popularly believed) the probability that the samples are from the same lot, but the probability of something else that is supposed to be indicative of this condition. The nature of this concept will become clear in the next few paragraphs. In order to quickly gather the necessary concepts for a logical consideration of the problem of significant differences, consider first the degree of probability rather than its nature, and for the moment suppose that one can find the probability that 2 samples under consideration came from the same lot.

If the probability were moderately remote and the decision not of great moment, one would be satisfied to assume a difference. If the decision were of a very serious nature, one would be inclined to demand that the probability be very, very remote, before being satisfied with assuming a difference. Hence, it is apparent that levels of significance are dependent upon engineering considerations, and a probability that would be considered as significant of a difference under one set of circumstances might not be considered significant under others. Therefore, choice of level of significance presents an economic problem which must be solved on engineering grounds.

Contrary to the assumption of the above paragraph, the probability that 2 samples came from the same lot presents a problem for solution which is an absolute philosophic impossibility. The truth of

this statement is easily seen if one considers the simple case of 2 samples by attributes, (c_1, n) and (c_2, n). It is an easy matter to find the probability that these 2 samples would be produced by any specific lot, Q_1. However, that may not be a satisfactory solution, for there is a just slightly different probability that they would be produced by a slightly different lot Q_2, etc.; and a moment's reflection will show that even the summation of the probabilities of all possible lots [2] would be without avail unless one knew the *frequency* of occurrence of lots of various Q. However, Q is in general unknown. Furthermore, the question of significant differences is most apt to arise when the samples are so small that even from the combined samples one cannot get a satisfactory estimate of Q. In fact, that is quite the crux of the matter. It therefore follows that instead of the probability that the samples came from the same lot, one must secure some alternative method of judging whether or not the samples appear to be from the same source.[3] The nature of the probability therefore raises a problem in estimation which must be solved on philosophic grounds.

The statistical method of solution is, of course, a problem in mathematical statistics, but no logical consideration is possible until at least partial solutions are obtained for the above two problems.

Thus, significant differences present three distinct problems: (*a*) degree of probability, (*b*) nature of probability, and (*c*) statistical method, which, in turn, must be considered on three bases: (*a*) economic grounds, (*b*) philosophic grounds, and (*c*) mathematical statistics.

A detailed theoretical discussion is undesirable in a brief presentation of practical usages. However, at least a brief discussion is absolutely necessary, (*a*) in order for the occasional user of statistical

[2] Solutions have been made on a basis of Bayes' theorem (see "Some General Results of Elementary Sampling Theory for Engineering Use," P. P. Coggins, *Bell System Technical Journal*, January, 1928). A very detailed solution for samples of attributes (with which the author disagrees) is given in *Methods of Estimating the Significance of Differences in or Probabilities of Fluctuations Due to Random Sampling*, G. F. McEwen, University of California Press, Berkeley, Calif., 1929. However, Appendix A indicates the grave danger of the assumption of a priori equal likelihood when samples are small; and, since the question of significance of differences is so often raised with respect to small samples, it is quite obvious that solutions on this basis may be most misleading.

[3] Note that the significance of differences answers a question of commonness of causation with respect to 2 samples of the same phenomenon. The question of commonness of causation with respect to 2 samples of different phenomena is a problem in correlation which is discussed in the next chapter.

methods to interpret his results properly and intelligently, (*b*) because there is such a widespread misconception to the effect that probabilities of sameness or difference are possible, and (*c*) because highly respected men in various fields of engineering, industry, and science for some quite inexplicable reason tend to abandon the good sense on which their reputations are founded, when faced with differences in samples, and rashly come to conclusions based on scant data without even a semblance of an investigation of the probable reality of the apparent difference. In this connection, it is not for a moment contended that a statistical test of significance is a sufficient criterion for action. Various practical considerations may be of equal or greater importance. It is merely contended that the statistical test of significance is quite too important a help to human judgment to be ignored with impunity.

SIGNIFICANT DIFFERENCES OF ATTRIBUTES—SMALL SAMPLES

As a basis of discussion, let it be supposed that 2 samples of 17 are tested and in one sample none proves defective and in the other sample 3 prove defective. Does this difference appear likely to be due to chance? At least one logical reason for doubting the sameness of lots sampled appears to inhere in the disparity of defectives; i.e., $3 - 0 = 3$, or, in more general terms, the disparity in ratios of defectives $3/17 - 0/17 = 3/17$; one might then inquire into the probability that a lot Q would produce this type of disparity in sampling results. Note that in this instance it is not the size of c, viz., 0 defectives out of 17 or 3 defectives out of 17, which raises the doubt (there would be little doubt if results were 0, 0 or 3, 3), but the disparity[4] $3/17 - 0/17$. Therefore, one does not wish to inquire into the probability that the lot Q would produce 0 defectives, 3 defectives, 3 or more defectives, but into the probability that the lot Q would produce an absolute (the sign does not matter[5]) difference in ratios as great as 3/17. Since probability runs on a scale from 0 to 1, and, since the disparity must be 0, 1/17, 2/17, $\cdots$ 17/17, it is better to consider the probability that the lot Q would produce a disparity in ratios as great or greater than 3/17.

[4] Contrast this situation with that in the control chart technique, where a number of samples were taken and a c which was greater than c_U or less than c_L raised a similar doubt. In that case, something was known about Q and the probability of various c's.

[5] If the sample sizes are different, a refinement can be introduced whereby the sign does become of some consequence.

This would appear to be a satisfactory approach to a solution of the philosophical problem of estimating a difference, if Q were known. It appears satisfactory for two reasons: first, it classifies the situation at hand into a category (that of finding a probability of a difference of ratios); second, the category is such that if this process is done repeatedly, and if a certain objective probability P_s is chosen as a rejection criterion, e.g., 0.01, as a level of significance, then in random sampling one will make the error of judging lots which are really the same to be different just $100P_s\%$ of the time.[6] That is, the method of estimation ties in with the level of significance in such a way that one can make an economic choice.

By way of factual nicety, one other point should be noted. As in all tests of this character, it is tacitly implied that sampling is random. That is to say, one makes no direct statement regarding the observed samples, but instead states that *if* samples are taken at random from a single lot of fraction defective Q, *then* a difference of so much will be observed such and such a percentage of the time. The statement is eternally true. *The cogency of its application to the samples in question, of course, inheres in the randomness of those samples.* However, in the practical case, the samples are not apt to be seriously biased,[7] and the experimenter himself is the best judge of the bias likely to exist in the type of sample under consideration. *Even if bias should exist, there is quite nothing that can be done about it.* In small samples, one cannot even test for randomness. Nevertheless, it would hardly be the part of wisdom to abandon the test in favor of arbitrary decision, because of inability to secure methodical assurance of randomness. As has been previously pointed out, the test is only an important guide to judgment; not a binding criterion for taking action contrary to that indicated by other plausible evidence. Hence, in this discussion, it will be assumed that sampling is random.

Calculation of the maximum probability of a chance difference. However, one does not and cannot know Q. A knowledge of Q is not always necessary. One knows that the standard deviation of the fraction defective is $(PQ/n)^{1/2}$. This is a maximum when $P = Q = 0.5$. Therefore, no lot is as likely to give any specific disparity [8] as

[6] It is quite impossible to say how frequently different lots will be erroneously judged the same. Obviously that depends upon how different the lots happen to be.

[7] See the discussion of randomness of sample at the beginning of Appendix A.

[8] More specifically, the standard deviation of the difference is $\sqrt{2}\,\sqrt{PQn}$, which is a maximum for $Q = 0.5$. See Chapter XII for a discussion of a variable which is the difference of two random variables.

the lot $Q = 0.5$. Therefore, if one finds the probability P_d that a lot of fraction defective 0.5 would give a disparity in 2 samples of 17 which is as great or greater than (3/17), then one knows that the probability associated with any other lot whatsoever is less than P_d. If P_d is less than the chosen level of significance, P_s, then there is no occasion for further inquiry. On these grounds, one would judge lots of the same Q to be different not more than $100P_s\%$ of the time. The method of calculating the particular case, $Q = 0.5$, is easy to see.

Before calculating this probability, it is well to reflect on just what it means. There are 18 possible values of c on the first sample, 0, 1, 2, $\cdots$ 17. A like number of values of c can occur on the second sample. Hence there are 18×18 pairs of samples possible (considering order). Of these possible pairs, (0,0), (1,1), (2,2), $\cdots$, (17,17) or 18 pairs yield a disparity of 0; (0,1) and (1,0), (1,2) and (2,1), $\cdots$, (16,17) and (17,16) or 34 pairs yield a disparity of 1/17; (0,2) and (2,0), (1,3) and (3,1), etc., or 32 pairs yield a disparity of 2/17, etc. One can calculate these respective probabilities for a specific case and sum them.[9] The actual summation of the probabilities indicated above for the case of $P(3/17)$ is 0.608. This is the probability of an absolute difference of less than 3/17. Therefore, $P_d = 0.392$. This is the probability of an absolute difference of 3/17 or greater. We know, therefore, *with absolute certainty* that the probability that any lot whatsoever would produce a disparity in ratios as great or greater than 3/17, in 2 random samples of 17 is equal to or less than 0.391. Unless we choose to assume that Q is much smaller than 0.5 (in which event P_d would be somewhat less than 0.39), this difference surely does not appear to be significant. It is possible that it could occur as frequently as 39 times in 100 with samples from the same lot.

It may be observed that the above calculation is very laborious.[10] However, one is not generally interested in the exact value of P_d,

[9] The general case of the probability of an absolute difference in ratios of less than x/n where $x/n = (c_1/n - c_2/n)$, can be written:

$$P\left(<\frac{x}{n}\right) = \sum_{c_1=0}^{n} P(c_1, c_1) + \sum_{d=1}^{x-1} \sum_{c_1=0}^{n-d} P(c_1, c_1 + d) + \sum_{d=1}^{x-1} \sum_{c_1=d}^{n} P(c_1, c_1 - d),$$

where $P(c_1, c_1 + d)$ is the probability of c_1 defectives on the first sample and $c_1 + d$ defectives on the second sample.

[10] The author has shown that the value of P to seven decimals can be read directly from the tables of the Incomplete Beta-Function Ratio by looking up $2I_{0.5}(n + x, n - x + 1)$ where $n + x = p$ and $n - x + 1 = q$ in the terminology of the tables. The proof of this relationship is tedious.

but only in knowing whether it is less than the chosen level of significance P_s. One can obtain this knowledge for $P_s = 0.2$ from Chart $0.1 = I_Q$ and for $P_s = 0.01$ from Chart $0.005 = I_Q$.

In the case of Chart $0.005 = I_Q$ find the value of c which corresponds to $Q = 0.5, n$. If c is greater than $n + x - 1$, the difference is not significant with respect to the 1% level of significance; i.e., $n + x - 1$ would have to be greater than c for the probability of the difference to be less than 0.01. The operation in the case of Chart $0.1 = I_Q$ is identically the same, but the interpretation is with respect to the 20% level of significance. This level of significance is too low to be of much practical use except for the purpose of showing how unreliable small samples are. This, however, is an important usage. The example of (0,17), (3,17) just cited was taken from actual experience. The experimenters, who were seasoned engineers, were so sure that the difference was significant that they went ahead with the experiment without awaiting the calculation of the significance of the difference, which at that time was being done by the method of summing the 84 probabilities. In the next 15 trials, the kind of article which gave 0 failures out of 17 actually gave 3 failures out of 15.

Calculation of likely probabilities of a chance difference—small samples of equal size. There will be times when one will not be satisfied with the maximum probability of a chance difference, especially if one has sound reasons for believing that true Q is very much less than 0.5. Under these conditions there are two plausible alternatives. It can be readily shown [11] that the most likely lot to produce two samples (c_1, n), (c_2, n) is the lot in which

$$Q = \frac{c_1 + c_2}{2n}.$$

If one chooses this value of Q (instead of 0.5) and sums the probabilities of appropriate pairs outlined above, then in repeated trials he will make the error of judging lots to be different, when they are the same, approximately $100P_s\%$ of the time. On the other hand, one may have reason for believing that neither of the samples is typical

[11] By principles previously given, the probability that the lot Q would produce the samples (c_1,n_1), (c_2,n_2) is:

$$P[(c_1,n_1), (c_2, n_2)] = \frac{n_1!}{c_1!(n_1 - c_1)!} Q^{c_1}(1 - Q)^{n_1 - c_1} \cdot \frac{n_2!}{c_2!(n_2 - c_2)!} Q^{c_2}(1 - Q)^{n_2 - c_2}.$$

To find the value of Q which makes $P[(c_1, n_1), (c_2, n_2)]$ a maximum, take the first derivative with respect to Q, equate to zero, and solve for Q.

of the kind of article under investigation, and wish to substitute some value of Q which from experience one knows is of the order of magnitude which should be encountered in the lots sampled. One generally has something of this type of knowledge independent of the sample. Under these conditions the frequency with which lots of the same fraction defective are judged to be different is a function of the excellence of the engineering judgment involved. It should be noted that a moderate error in the solution of Q introduces only an error of secondary order in P_d.

For the more general case of Q not equal to 0.5, it appears desirable to derive a general procedure, even though it may be less precise than the one given for $Q = 0.5$. It may be observed that the difference, x/n, where $c_1/n - c_2/n = x/n$, is symmetrically distributed about zero as a mean irrespective of the value of Q. This is true because every x/n really bears a $\pm$ sign, the sign merely depending upon which ratio is subtracted from the other. Under the hypothesis that the lots are the same, it is also known that the modal value of x/n is zero. Therefore, it is known that x/n approximates the normal curve, is symmetric, and unimodal, with mode and mean at zero and standard deviation [12] equal to $\sqrt{2}\sqrt{PQ/n}$. The probability of a x/n less in absolute value than that observed can therefore be approximately [13] summed by means of the normal law. $P(<x/n)$ may therefore be found by looking up the probability associated with the range

$$t = \frac{(x/n) - (1/2n)}{\sqrt{2}\ \sqrt{PQ/n}}$$

This value of P can be found by reading $t = x/\sigma$ on Fig 10·4. However, the actual probability is rather unimportant, and the mere observation that t is greater than 2 (for approximately a 5% level of significance) or 3 (for a very cogent level of significance) is generally quite enough, for obviously no great importance can be attached to exact probabilities where Q is inferred from a small sample. It should be noted that for small sample sizes the subtraction of $1/2n$ from

[12] See Chapter XII.

[13] For Q ranging from 0.10 to 0.5, and $n = 10$, the approximation is correct to -0.003, for the working range $P = 0.05$ to $P = 0.005$. For very small Q, the method is worthless; likewise, small samples are worthless for detecting a difference in such a range. When P is large, the answer may be in error by several per cent; however, when P is large, no significant difference is indicated. It may therefore be observed that the method is a good working approximation in the practical range.

x/n is important. In this instance one is approximating a discontinuous function (one that moves by jumps) by a continuous function. Therefore the cutoff point on the continuous function is selected midway between the jumps. The jumps can occur only at 0, $1/n$, $2/n$, etc., whence the midpoints are at $1/2n$. Hence $1/2n$ is subtracted from the range which is being summed.

Returning to the example $c_1/n - c_2/n = |0/17 - 3/17| = 3/17$, let us test for significance, using the most likely Q, viz., $Q = (0 + 3)/2n = 0.088$. In this case, $x/n = 3/17$; therefore

$$t = \frac{3/17 - 1/34}{\sqrt{2}\sqrt{(0.088)(0.912)/17}} = 1.52.$$

From Fig. 10·4 an x/σ of 1.52 gives an area to the left of the ordinate of 93.4%. The area included in the symmetric range of $\pm t$ is $2(0.934 - 0.50) = 0.868 = P(x/n)$. Therefore $P_d = 0.132$. This is decidedly less than the probability of 0.392, which was the greatest probability that any lot could have of producing an absolute difference as great or greater than 3/17, but surely not significant at that. It should be further noted that significance, in the case of assuming a small value of Q, is greatly lessened by virtue of the assumption; i.e., lots of small Q are less likely to produce differences.

Let us use this approximate method for the case of $Q = 0.5$ where it is known that $P_d = 0.392$.

Here,

$$t = \frac{3/17 - 1/34}{\sqrt{2}\sqrt{(0.5)(0.5)/17}} = 0.858, \text{ and } P_d = 0.39.$$

The approximation is surely satisfactory. This is as expected since $P = Q = 0.5$ and a fairly large n have rendered the distribution almost normal.

Significant differences of attributes for samples of different sizes. It sometimes happens that one wishes to test for a significant difference in samples which are not the same size. Consider the general case c_1/n_1, c_2/n_2 (let it be assumed that n_2 is greater than n_1 but that the subscripts have no meaning of magnitude with reference to the c's). In this case, the approximate method just described still applies, but some care must be exercised in the selection of the cutoff point. By way of distinction, let

$$\frac{c_1}{n_1} - \frac{c_2}{n_2} = \frac{n_2c_1 - n_1c_2}{n_1n_2} = \frac{\Delta}{n_1n_2}.$$

Then,

$$t = \frac{(\Delta/n_1n_2) - [(\Delta - \Delta_i)/2n_1n_2]}{\sqrt{PQ(1/n_1 + 1/n_2)}},$$

where Δ_i is the next Δ lower than the observed Δ which can exist. The selection of Δ_i must be made by trial but is not difficult. Subtract 1 from Δ and try this new Δ in the expression;

$$\frac{\Delta_i + kn_2}{n_1} = \text{an integer},$$

where k is allowed to take successive values 0, 1, 2, etc., but subject to the restriction that $\Delta_i + kn_2$ shall not be greater than n_1n_2.

As an extreme example consider the samples 3/3, 3/17. In this case

$$\frac{c_1}{n_1} - \frac{c_2}{n_2} = \frac{3}{3} - \frac{3}{17} = \frac{51 - 9}{51} = \frac{\Delta}{51} = \frac{42}{51}$$

Find the next lower jump to 42/51.

Allow $\Delta_i = 41$.
If $k = 0$, then $\frac{41 + 0}{3} \neq$ integer.
If $k = 1$, then $41 + 17 > 51$.
Allow $\Delta_i = 40$.
If $k = 0$, then $\frac{40 + 0}{3} \neq$ integer.
If $k = 1$, then $40 + 17 > 51$.
Allow $\Delta_i = 39$.
If $k = 0$, then $\frac{39 + 0}{3} = 13$.

Therefore 39/51 is the next lower jump. Taking the most likely value of Q, viz., $(3 + 3)/(3 + 17)$,

$$t = \frac{\frac{42}{51} - \frac{42 - 39}{102}}{\sqrt{(0.30)(0.70)(\frac{1}{3} + \frac{1}{17})}} = 2.77.$$

Therefore $P_d = 0.0058$ approximately. The value of P_d computed from 7-figure tables of the Incomplete Beta-Function Ratio is 0.0054574.

For small Q or large probabilities and the samples unequal, the approximation is not so close. In Table 11·1 for ($n_1 = 3$, $n_2 = 17$)

the approximate value of P_d is shown on top, the exact value on the bottom, and the most likely Q is used in all cases.

TABLE 11·1

COMPARISON OF EXACT AND APPROXIMATE PROBABILITIES OF THE SAMPLES $(c_1,3)$ AND $(c_2,17)$

Δ \ Q	0.05	0.10	0.20	0.30	0.40
17/51	0.0176 0.0639548				
18/51				0.2187 0.2783615	
24/51					0.1416 0.1493336
28/51			0.0316 0.0378888		
34/51		0.0005 0.0054982			
36/51					0.0253 0.0191799
42/51				0.0058 0.0054574	
48/51			0.00026 0.0009458		

If n_1 differs but little from n_2, one had best try the method for both n's equal and see if it appears to be necessary to use the more complicated method for unequal n's.

SIGNIFICANT DIFFERENCES OF ATTRIBUTES—LARGE SAMPLES

If samples are large the approximation methods given for small samples not only hold but also, of course, are more accurate. In fact, if the n's are of the order of 50 or greater, and Q is not very small, one may dispense with the refinement of splitting the difference between the jumps, and write

$$t = \frac{(c_1/n_1) - (c_2/n_2)}{\sqrt{PQ(1/n_1 + 1/n_2)}},$$

where Q is taken as $(c_1 + c_2)/(n_1 + n_2)$. If samples are large, observed ratios should not differ greatly from true ratios; and therefore it would hardly be plausible to assume Q to be 0.5, when observed ratios are considerably different from this value.

If n_1 is very large and n_2 much smaller, the question of the significance of the difference can often be answered by observing the ratio

$$t = \frac{(c_1/n_1) - (c_2/n_2)}{\sqrt{}}$$

where Q is inferred from n_1 (the large sample) and the n under the radical is taken from the other [14] (the small sample). If the ratio is greater than 3, the samples appear to be from different cause systems; if the ratio is less than 2, there appears to be no statistical reason for suspecting a difference. For large samples (neither sample less than 50) the chi-square test is more elegant. A simple explanation of this test is given in Chapter XXII of Yule and Kendall's book.

SIGNIFICANT DIFFERENCES OF VARIABLES

In sampling by variables one is faced with all the problems mentioned in the general discussion at the beginning of this chapter. Furthermore, instead of merely wishing to know the unknowable about Q, one now wishes to know the unknowable about at least two statistics, $\bar{X}$ and σ. Statisticians have therefore tried to devise tests which are predicated on distributions which are either independent of these parameters or dependent on only one of them. Two well-known distributions of this type are Student's t-distribution and R. A. Fisher's z-distribution. The former is used in connection with means; the latter in comparing two standard deviations. These distributions have been widely advocated in connection with biological [15] and agricultural work,[16] but simpler methods appear to find favor in industrial work. The test for neither statistic is entirely free of limitations on the other statistic; and in application each involves

[14] Note the difference in philosophy in this case and also its cogency. Let one infer Q from (c_1, n_1). Then, if n_1 is large, Q must be near the true value and c_1/n_1 near the true average value. The ratio then measures in standard deviations the probability of a value as far away from the average as $c_1/n_1 - c_2/n_2$, which is a most dependable measure of significance in this case.

[15] *Statistical Methods for Research Workers*, R. A. Fisher, Oliver and Boyd, London, 1936.

[16] Chapters 3 and 10, *Statistical Methods*, George W. Snedecor, Collegiate Press, Ames, Iowa, 1938.

the assumption of normality. In the application of statistical methods, one must consider the extent to which one wishes to enter into refinements, the concomitant necessity of appraising the real and implied assumptions upon which the method is predicated, and the logic underlying the inferences drawn therefrom. In the light of this cautionary statement, two simple and rugged tests are recommended to the occasional user of statistical methods.[17]

Two averages appear to be significantly different if

$$\frac{|\bar{\bar{X}}_1 - \bar{\bar{X}}_2|}{\sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}}} > 3.$$

Two standard deviations appear to be significantly different if [18]

$$\frac{|\sigma_1 - \sigma_2|}{\sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2}}} > 3.$$

In these formulas $\bar{\bar{X}}_1$ and $\bar{\bar{X}}_2$ are the calculated values of the average, unless one of them is known or assumed; and $\sigma_{\bar{X}_1}$, $\sigma_{\bar{X}_2}$, σ_{σ_1}, and σ_{σ_2} are calculated values of the standard deviation of the average and standard deviation of the standard deviation, respectively, unless one or both are known or assumed. These formulas are valid subject only to the restrictions that the population or populations sampled are controlled (in the control chart sense) and that the samples are not very small. It may be recalled that, for a sample [19] of n, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ and $\sigma_{\sigma} = \frac{\sigma}{\sqrt{2n}}$.

The ratio of the difference of two means to the standard deviation of their difference would be normally distributed [20] if true values of σ were used instead of estimated values and if the parent population

[17] For a very readable explanation of more precise tests of significance, see Croxton and Cowden, *Applied General Statistics*, Chapters XII and XIII, Prentice-Hall, New York, 1939.

[18] These tests consist merely of dividing a difference by the standard deviation of that difference. In general terms, two observations of any statistic θ appear to be significantly different if

$$\frac{|\theta_1 - \theta_2|}{\sqrt{\sigma^2_{\theta_1} + \sigma^2_{\theta_2}}} > 3$$

for the simple reason that deviations in excess of 3 standard deviation seldom occur.

[19] For the standard deviation of a standard deviation computed from k subgroups of m (where $km = n$), see Chapter XII.

[20] Note comments of Chapter V on "Distributions."

were normal. With a controlled parent population (not necessarily a normal one) the ratio is still almost normally distributed if the samples are as large as 5. The experimenter himself is the best judge of the accuracy of the estimates of σ. Thus, it is apparent that, whereas no numerical probability can be attached to this test, its use as a criterion will seldom lead to the error of judging lots which are the same to be different, even if samples of the order of size 10 are used. In samples of this size a test of control could not be applied, and the state of control would have to be inferred from independent knowledge.

In the inequality for standard deviations, the ratio of the difference of two standard deviations to the standard deviation of their difference would not be normally distributed even if the correct value of σ (as represented by the radical) were available and if the parent population were normal. This is true because the standard deviations of samples of n from a normally distributed population are not themselves normally distributed, but only approach normality with large n. However, for sample sizes of 5 or greater, from a normal universe, one may safely assume that the Camp-Meidell inequality applies.[21] However, if the parent universe departs appreciably from normal, statistical theory cannot in general predict the distribution of the standard deviation, and one can safely assume only that the Tchebycheff inequality applies. This very melancholy estimate does not include a consideration of the fact that the radical in the denominator of the inequality is an estimate of the standard deviation of the difference rather than the standard deviation itself. It is thus apparent that the level of significance represented by this inequality may be as poor as 0.10.

However, it is not recommended that it be improved by substituting a 4 for the 3, as such action would materially lessen the probability of detecting the differences when they exist. With very small sample size, no test is dependable unless the difference is overwhelming; with sample sizes of the order of 20 from controlled universes the test will give moderately satisfactory results in practice; with larger sample sizes taken under controlled conditions, results, of course, become more dependable.

Illustration of significant differences of variables. Suppose that two types of bullets have been fired under essentially the same conditions with results as shown in Table 11·2. Only the average, $\bar{X}$, and observed standard deviation, σ, of the chronologically arranged

[21] Shewhart, *op. cit.*, Chapter XIV.

groups of 5 are shown. The question is raised whether these two types of bullets appear to be significantly different (*a*) with respect to aver-

TABLE 11·2

COMPARISON OF VELOCITY CHARACTERISTICS OF TWO TYPES OF BULLETS

22 groups of 5 for bullet type 1		20 groups of 5 for bullet type 2	
σ	$\bar{X}$	σ	$\bar{X}$
75	2,900	145	2,805
52	2,855	50	2,770
172	2,715	65	2,550
244	2,820	160	2,660
75	2,790	130	2,705
152	2,855	60	2,745
177	2,640	215	2,660
85	2,955	150	2,525
70	2,860	35	2,650
167	2,780	55	2,720
200	2,730	175	2,530
125	2,865	105	2,810
40	2,685	40	2,605
100	2,870	80	2,680
110	2,845	85	2,595
112	2,770	90	2,720
138	2,775	120	2,780
70	2,665	70	2,665
120	2,820	100	2,725
60	2,960	200	2,700
25	2,805		
80	2,750	Total 2110	Total 53,600
Total 2649	Total 61,710		
Aver. 120.41	Aver. 2,805	Aver. 105.5	Aver. 2,680

$$\text{Est. } \sigma_1 = \frac{120.41}{0.841} = 143.3.$$

$$\sigma_{\bar{x}_1} = \frac{143.3}{\sqrt{5}} = 64.0.$$

$$\sigma_{\sigma_1} = \frac{143.3}{\sqrt{10}} = 45.2.$$

$$\text{Est. } \sigma_2 = \frac{105.5}{0.841} = 125.5.$$

$$\sigma_{\bar{x}_2} = \frac{125.5}{\sqrt{5}} = 56.2.$$

$$\sigma_{\sigma_2} = \frac{125.5}{\sqrt{10}} = 39.6.$$

age velocity (level of the quality characteristic), and (*b*) with respect to velocity dispersion (uniformity of the quality characteristic).

As a first step, one should test to see whether or not a state of statistical control appears to exist, as no predictions of the usual type are possible in the absence of statistical uniformity. Therefore, graphical control charts [22] as shown in Figs. 11·1*A* and 11·1*B* should be constructed. The estimated σ' and control limits can be read directly from charts as described in Chapters V and VI, although the calculation is indicated in Table 11·2. No points fall outside of the limits, so there is no reason for believing that the universes in their present state are not statistically controlled.

Apply the criterion for a significant difference in averages:

$$\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}}} > 3;$$

that is to say, one makes the hypothesis that there is no significant difference in average, and then tests to see whether observed results would be likely under the terms of this hypothesis. The first average was calculated as 110 observations; therefore $\sigma_{\bar{X}_1}$ = estimated σ'_1 divided by $\sqrt{110}$. Hence,

$$\sigma^2_{\bar{X}_1} = \frac{(143.3)^2}{110}.$$

In like manner,

$$\sigma^2_{\bar{X}_2} = \frac{(125.5)^2}{100}.$$

Hence, one raises the question

$$\frac{2805 - 2680}{\sqrt{\dfrac{(143.3)^2}{110} + \dfrac{(125.5)^2}{100}}} > 3?$$

$$\frac{125}{15.65} > 3.$$

The ratio of the difference to the standard deviation of the difference is 8. It is practically certain, therefore, that bullet type 1 has a significantly greater velocity than bullet type 2, as a deviation of 8 standard deviations can scarcely be attributable to chance causes.

[22] Since the samples are arranged in chronological order of firing, this process tests for control in the testing process. It does not necessarily test for control in the universes sampled, since the universe or lot may have been thoroughly mixed, i.e., order thereby lost, prior to the testing process. An uncontrolled product, if thoroughly mixed and then sampled, may not show a lack of control. Hence, in control chart technique, order is an important consideration. If order is lost, there is nothing that can be done about it, since by no known process can one restore it.

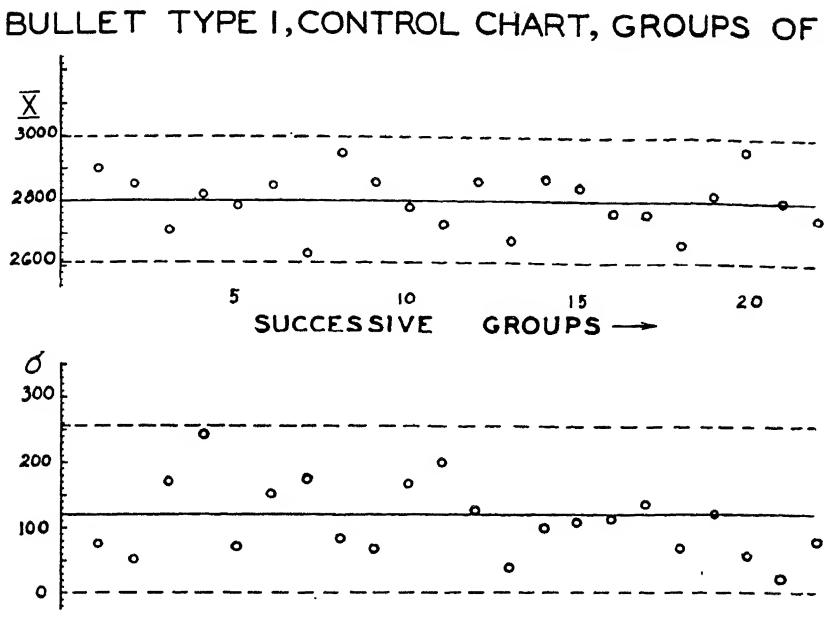


FIG. 11·1A.

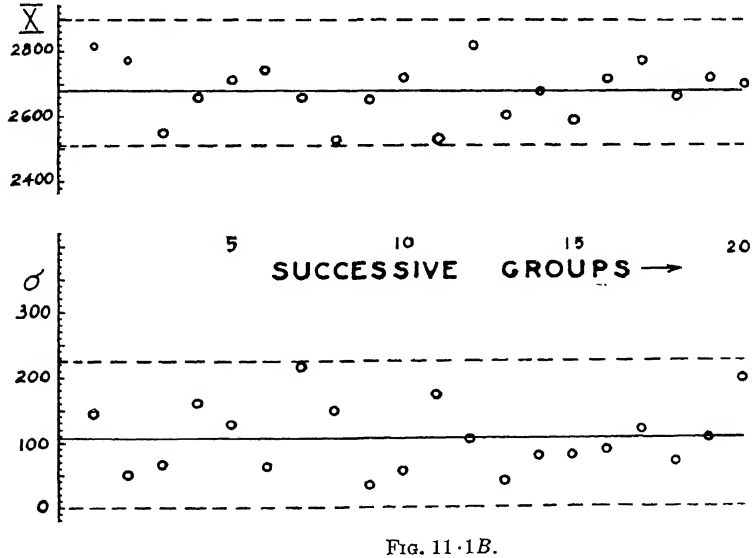


FIG. 11·1B.

One must be cautious in applying a like test to the standard deviations. The first standard deviation was not calculated directly from 110 observations, but from 22 subgroups of 5; therefore σ_{σ_1} is not equal to $143.3/\sqrt{2(110)}$. Of course, one could go back and calculate σ over the whole 110 observations (which were not actually tabulated, in the present case), and this procedure would be the most accurate. However, the σ from k subgroups of m is generally available, since it is obtained as a by-product of testing for control; and, although it is less precise (has a larger standard deviation of the standard deviation) than the σ calculated from the n observations, it has the same expected value, and one will generally use it for the sake of convenience. It is suggested, therefore, that the reader turn momentarily to Fig. 12·1, and enter the chart with $m = 5$ at the bottom and read the abscissa corresponding to the line marked $k = 22$. The reading is 0.0775. A universe of unit standard deviation would have $\sigma_\sigma = 0.0775$. Therefore, $\sigma_{\sigma_1} = 0.0775(143.3) = 11.12$. In like manner $\sigma_{\sigma_2} = 8.12$. Therefore

$$\frac{143.3 - 125.5}{\sqrt{(11.12)^2 + (10.20)^2}} > 3?$$

$$\frac{17.8}{15.15} \ngtr 3.$$

Therefore, upon the evidence submitted there is no reason for believing that the two types of bullets are significantly different in standard deviation.

One would therefore conclude: (*a*) that bullet type 1 has a significantly higher muzzle velocity and that the difference appears to be of the order of 100 feet per second; (*b*) that bullet type 1 may be less uniform in muzzle velocity but that the difference, if any, in variability is not significant as judged from this test.

Under these circumstances one might be tempted to resort to calculating the sigmas from the n observations. However, again looking at the chart for precision of estimates of σ, one sees that σ_σ or 20 subgroups of 5 is 0.0812; σ_σ for 1 group of 100 is 0.0711. The reduction in the standard deviation of the difference would therefore be of the order of only 12%. This could scarcely be expected to change the ratio 17.8/15.15 to a ratio greater than 3. Therefore such effort is unwarranted.

CHAPTER XII

MISCELLANEOUS STATISTICAL TECHNIQUES

Statistics? I can prove anything
by statistics except the truth.
—GEORGE CANNING (1770–1827)

Comments on measures of dispersion. In the brief discussion of the concept of frequency distribution which was given in Chapter V, the standard deviation was used as the measure of dispersion. Mention was also made of the probable error ($0.6745\,\sigma$), which is applicable when the distribution is strictly normal; and of the simpler, but less precise, statistic range, R. In order to provide a minimum of necessary tools for ordinary statistical analyses, this knowledge should be supplemented with a brief discussion of the relative merits of and the relationships between several measures of dispersion.

The standard deviation. The standard deviation was defined as

$$\sigma = \sqrt{\frac{\sum_{i=1}(\overline{X} - X_i)^2}{n}}$$

and it was pointed out that the standard deviation in small samples is usually smaller than the true standard deviation of the universe. In order to overcome this tendency (on the average) a correction factor, c_2, was introduced for use in estimating the universe σ'; but, in the control chart technique as outlined, the correction was automatically provided by the charts for averages and standard deviations. It is sometimes desirable to use these factors [1] directly. A list is given in Table 12·1.

[1] These are merely solutions of the equation:

$$\overline{\sigma} = \sqrt{\frac{2}{n}}\,\frac{\left(\frac{n-2}{2}\right)!}{\left(\frac{n-3}{2}\right)!}\,\sigma' = c_2\sigma',$$

which follows from the distribution function of the standard deviation published by "Student," *Biometrika*, Vol. VI, 1908. Actually, the so-called Student distribution was discovered much earlier by Helmert. F. R. Helmert, *Astronomische Nachrichten*, Vol. 88, No. 2096, 122 (1876).

TABLE 12·1

CORRECTION FACTORS c_2 FOR σ

n	c_2	n	c_2	n	c_2	n	c_2
		11	0.9300	21	0.9638	55	0.9863
2	0.5642	12	0.9359	22	0.9655	60	0.9874
3	0.7236	13	0.9410	23	0.9670	65	0.9884
4	0.7979	14	0.9453	24	0.9684	70	0.9892
5	0.8407	15	0.9490	25	0.9697	75	0.9900
6	0.8686	16	0.9523	30	0.9748	80	0.9906
7	0.8882	17	0.9551	35	0.9784	85	0.9912
8	0.9027	18	0.9577	40	0.9811	90	0.9916
9	0.9139	19	0.9599	45	0.9832	95	0.9921
10	0.9227	20	0.9619	50	0.9849	100	0.9925

The standard deviation of the standard deviation. Consider 100 measurements of a variable $X_1, X_2, \cdots, X_{100}$. Suppose that it is desirable to obtain the standard deviation of these 100 measurements, and also to estimate the standard deviation of the standard deviation. One method of procedure consists of taking the root mean square over all the measurements; i.e.,

$$\text{est. } \sigma' = \frac{1}{c_2}\sqrt{\frac{\sum_{i=1}^{n}(\bar{X} - X_i)^2}{100}}.$$

In this event, $\sigma_{\sigma'} = \sigma'/\sqrt{200}$. As a basis of discussion, suppose that $\sigma' = 1$; then $\sigma_{\sigma'} = 0.0707$.

However, it is laborious to take the root mean square over the whole number of observations. Work would be greatly reduced by breaking the data into 10 subgroups of 10; i.e., k subgroups of m, where $n = km$. In this event, est. $\sigma' = \bar{\sigma}_m/c_2$, where

$$\sigma_m = \sqrt{\sum_{i=1}^{m}\frac{(\bar{X} - X_i)^2}{m}},$$

and

$$\bar{\sigma}_m = \frac{1}{k}\sum_{j=1}^{k}\sigma_j.$$

Since, as previously pointed out, these two estimates of σ' have the same expected value, assume for illustration that $\sigma' = 1$. How-

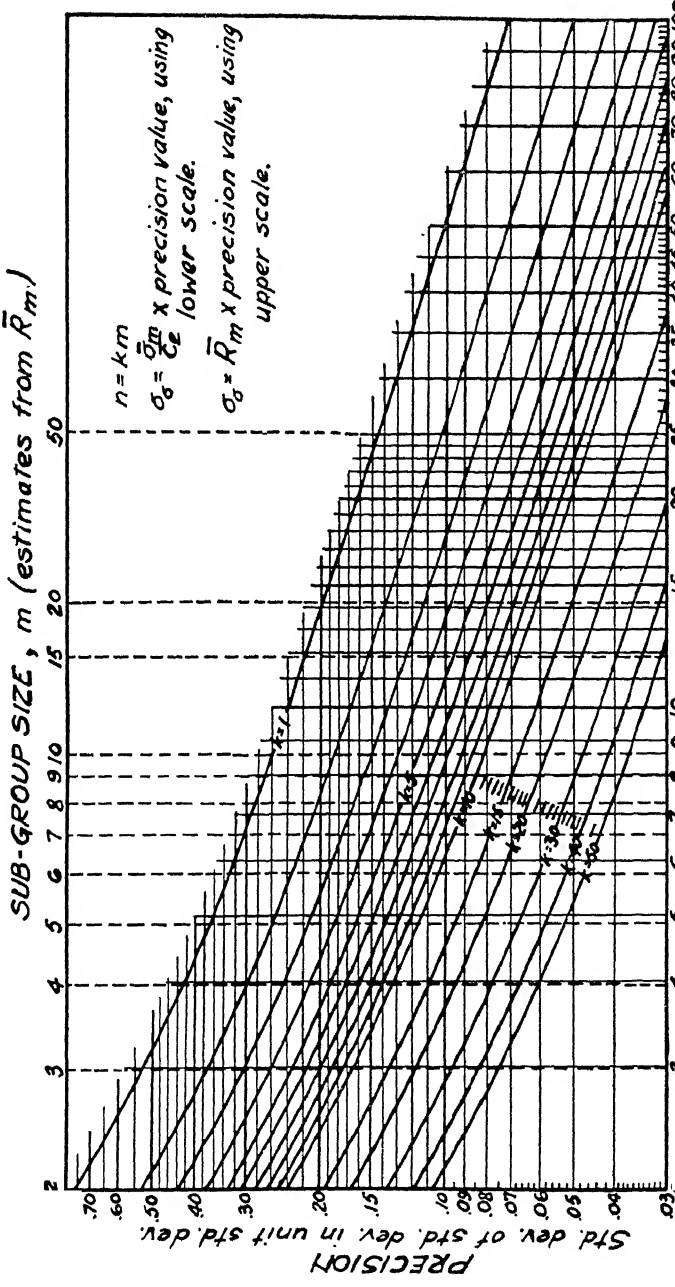

PRECISION OF ESTIMATES OF σ
σσ resulting from breaking sample of n into k sub-groups of m, and from use of range instead of std. dev.
SUB-GROUP SIZE, m (estimates from R̄m)
n = km
σσ = c̄m/c̄e x precision value, using lower scale.
σσ = R̄m x precision value, using upper scale.
PRECISION
Std. dev. of std. dev. in unit std. dev.
SUBGROUP SIZE, m (estimates from σm)

FIG. 12·1.

ever, σ_σ is not in this case $\sigma'/\sqrt{200}$, because the n used was not 100; it was 10 groups of 10, which is quite a different matter. Under these circumstances the expression for σ_σ does not reduce to simple form, but Fig. 12·1 provides an easy means not only for reading standard deviations of the standard deviation but also for comparing the relative precision of different methods of grouping with a view to choosing the most economic method.[2] In the present example, the intersection of the curves for $k = 10$ and $m = 10$ (reading the bottom scale) is opposite $\sigma_\sigma = 0.0745$. The loss in precision is slight and (as can be seen by noting the value of m corresponding to 0.0745 on the $k = 1$ curve) is the same as the precision which would be obtained from a single group of 92. In like manner, one can see that σ_σ for 20 samples of 5 is 0.081; and, finally, for 50 samples of 2, σ_σ is 0.107. This is serious, as it corresponds to a loss of more than 50% in efficiency. Had estimated σ' been some value other than unity, e.g., 176.42, then σ_σ for 10 samples of 10 would be $0.0745 \times 176.42 = 13.14$. That is, σ_σ can be found from Fig. 12.1 for any σ, if k and n are known.

This method of getting σ_σ is often very convenient, as data are often in the form of rational subgroups for control chart purposes (as such a test is so frequently desired before applying any statistical measures). Therefore, $\bar{\sigma}_m$ may be readily available, whereas the calculation of estimated σ' from the whole sample would be a new, laborious, and perhaps not merited operation.

It should be noted, however, that, when a single small sample comprises all the data, there is no point in multiplying the observed standard deviation of a small sample by $1/c_2$, as the observation itself is of little merit. The operation of the factor c_2 is of particular merit in connection with control chart technique. For general purposes it is to be depreciated as it does not readily lend itself to algebraic use. In the remainder of this chapter the use of the factor c_2 should not be implied where it is not specifically stated.

Variance. Variance is defined as

$$\sigma^2 = \sum_{i=1}^{n} \frac{(\bar{X} - X_i)^2}{n}.$$

[2] Figure 12·1, of course, is predicated on a normal universe. This assumption will be implied throughout this chapter, except where otherwise noted. Knowledge of the relationships between statistics of universes other than normal is very scant, although some of the relationships of the normal universe appear to apply fairly well to universes that are approximately normal.

The estimated universe variance, which is

$$\frac{n}{n-1}\sigma^2,$$

does not suffer from the algebraic disadvantages associated with σ and is somewhat superior from the statistical viewpoint. From the engineering viewpoint the interpretation of a non-linear measure is awkward. For example, it is difficult to picture a case where practically all observations fall within $\pm t\sigma^2$ as one does in the case of $\pm 3\sigma$. However, no such use is ordinarily made of variance.

Much is to be said for a standard deviation defined as the square root of estimated universe variance.[3] Although its distribution is different from standard deviation as defined by

$$\sigma' = \sqrt{\sum \frac{(\bar{X} - X)^2}{n}},$$

it not only is algebraically more tractable but also approaches σ' more rapidly, and has a smaller standard deviation. Although it lacks a background of successful usage in control chart work, like that associated with standard deviation, there is certainly good reason to believe that it would work satisfactorily in a slightly altered control chart technique, and it may be a development of the future.

Range (bracket or maximum dispersion). The simplest of all statistics, the one which is intuitively a part of man's natural mode of thinking and one of quite popular use, is the mere difference between the greatest and least values in a group. That is to say, having made n observations of a variable, $X_1, X_2, \cdots, X_n$, if the range $(X_{\text{max.}} - X_{\text{min.}})$ is very small, one naturally concludes that a high degree of uniformity exists. The interpretation of range, however, is very dependent upon the sample size; e.g., a range of 0.14 second in a sample of 15 anti-aircraft time fuzes at a 10-second setting is much less disturbing than the same range of 0.14 second on a sample of 5 fuzes at the same setting. It is therefore desirable in comparing

[3] Chapter 23 of Yule and Kendall contains a lucid explanation of this definition of σ. Briefly σ is estimated from $\bar{X} - X$, where $\bar{X}$ is the observed mean instead of the true mean. This unknown parameter results in the loss of one degree of freedom; hence $n - 1$. That is, having availed oneself of the $\bar{X}$, one could find any X, given the $n - 1$ remaining X's. In like manner, when work involves several parameters which are estimated from data, one generally loses one degree of freedom for every parameter so estimated.

ranges to reduce them to some common denominator. A convenient method of accomplishing this end consists of converting the observed range to the universe standard deviation which would (on the average) produce such a range. This can be accomplished by multiplying the observed range by a factor, just as one multiplies the observed standard deviation by the factor $1/c_2$. By way of analogy, this factor is noted herein as $1/d_2$. Table 12·2 gives the factors [4] d_2 for $n = 2$ to $n = 15$. It should be noted that d_2 is merely the average number of standard deviations included in a bracket of a sample of n. Naturally, d_2 for n very large is slightly greater than 6 (it reaches this value at $n = 444$); i.e., $\pm 3\sigma$ includes practically the range of the universe. There is very little point, except for purposes of comparison, in multiplying an observed range of a small sample by $1/d_2$, just as there is little point in multiplying an observed

TABLE 12·2

FACTOR d_2 FOR CHANGING R TO EST. σ'

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
d_2	1.128	1.693	2.059	2.326	2.534	2.704	2.847	2.970	3.078	3.173	3.258	3.336	3.407	3.472

standard deviation of a small sample by $1/c_2$. The observation itself is quite too indefinite to merit a meticulous operation. However, there is a great deal of point in multiplying the average range, $\bar{R}$, of k samples of size m by $1/d_2$. This gives a very good estimate of σ' and is exceedingly easy to compute.

The standard deviation of the standard deviation (estimated from range). Suppose that est. $\sigma' = (1/d_2)\bar{R}_m$, where $\bar{R}_m$ is the mean range [5] of 10 samples of 10. As in the case of estimated σ' from k samples of n, $\sigma_{\sigma'}$ is not equal to $\sigma'/\sqrt{2n}$. Figure 12·1 gives $\sigma_{\sigma'}$ for a wide range of k and m, by reading m on the top scale. Thus, if est. σ'

[4] The expected value and standard deviation of the distribution of ranges of samples of size n from a normal universe is known. See L. H. C. Tippett, "On the Extreme Individuals and the Range of Samples Taken from a Normal Population," *Biometrika*, Vol. XVII, 1925. The relationship between range and probable error for a normal universe was also derived independently in an unpublished paper by Dr. L. S. Dederick of the Ballistic Research Laboratory, in connection with work in gunfire.

[5] See O. L. Davies and E. S. Pearson, "Methods of Estimating from Samples the Population Standard Deviation," *Supplement to the Journal of the Royal Statistical Society* (Industrial and Agricultural Research Section), Vol. I, No. 1, London, 1934.

for 10 samples of 10 is 1, σ_σ is found to be 0.081 (note that this is slightly larger than σ_σ for 10 samples of 10 using standard deviation, which was 0.0745). In like manner, σ_σ for 20 samples of 5 (est. σ' being unity) is 0.082, which is only very slightly inferior to that obtained with standard deviation. It is thus seen that, by an examination of Fig. 12·1, one can inquire into the relative merits of various choices of k and m and of range and standard deviation.

Comments on range. Writers on statistics are almost unanimous in their condemnation of range as the worst possible measure of dispersion. Its marked loss in efficiency as n increases is all too evident from Fig. 12·1. For example, a sample of 25, using standard deviation, yields precision equal to a sample of 50, using range. However, as the sample size decreases, the disparity in efficiency also decreases; and, of course, the statistics are equally efficient for $n = 2$. For small samples ($n = 2$ to $n = 10$ or 15) the loss in efficiency is relatively slight. If sampling is being done by small subgroups, as is general in quality control work, it may be easier and more economic to take an additional measurement and use the simpler statistic. In the destructive sampling of expensive articles of ammunition, the use of range, even for small subgroups, is not recommended, for even a loss of 4 or 5% in efficiency means a substantial monetary loss. However, in a system for process inspection by the control chart technique as described in Chapter VII, the use of range instead of standard deviation has been found to be not only entirely satisfactory from the engineering and statistical point of view, but also valuable from the practical point of view, because it makes the institution of quality control (with its many economic advantages) possible without the necessity of either changing plant personnel or giving existing personnel any special training. This subject is discussed further in Appendix C.

Mean deviation. The mean deviation is historically of importance and still enjoys considerable currency. It is defined as

$$\text{M.D.} = \frac{\Sigma\,|\,\bar{X} - X\,|}{n}$$

With appropriate procedure, it can be used for almost all the purposes for which other measures of dispersion are used. The average mean deviation of k samples of n can be changed to estimated standard deviation by multiplying the average mean deviation by $1.253\sqrt{n/(n-1)}$. The use of mean deviation, however, is not

recommended for two cogent reasons: it has (*a*) neither the efficiency [6] of the standard deviation, (*b*) nor the simplicity of range.

By way of protection from the wrong use of statistics by others, one or two dangers associated with mean deviation should be pointed out. For small samples, the mean deviation is very sensitive to sample size; and, unless the sample size is quoted, mean deviation is almost meaningless. This difficulty can be overcome by multiplying the observed M.D. by $\sqrt{n/(n-1)}$. However, many people who use mean deviation are aware neither of the necessity for the correction nor of the manner of accomplishing it.

In older work it was popular practice to multiply mean deviation by 0.8453 (known as Peter's formula) to estimate probable error, in order to avoid the labor of getting the root mean square deviation (standard deviation), which would then have to be multiplied by 0.6745. This procedure was predicated on the principle that, for large n, $\sigma = 1.253$ M.D. and $0.6745 \times 1.253 = 0.8453$. There are two objections to Peter's formula in addition to those already cited with respect to probable error: (*a*) the loss of efficiency [7] due to estimating the standard deviation from the first moment; (*b*) failure to take account of sample size.

Occasionally one encounters work in which a mean of deviations is taken, not from the mean of the observations themselves, but from some arbitrary or aimed-at value. This mean of the deviations from the arbitrary value is then treated as a mean deviation. Such procedure is too fallacious to bear comment. Strange to say, this procedure, under certain circumstances, has given sufficiently satisfactory approximations for its error not to be obvious through its false results.

Successive differences. A method has been developed for estimating the standard deviation from the mean square successive difference δ^2, where

$$\delta^2 = \frac{\sum_{i=1}^{n-1}(X_i - X_{i+1})^2}{n-1}$$

[6] For normal populations, no other measure of dispersion can be as efficient as that calculated from the second moment; i.e., the standard deviation and variance. For an interesting proof, see Chapter VII, *The Calculus of Observations*, E. T. Whittaker and G. Robinson, Blackie and Son, London, 1929.

[7] For large n (greater than 25), mean deviation is 88% efficient as compared with standard deviation. For small n (2 to 10), it is comparable to range.

Let the symbol $\sigma(\delta)$ designate the estimate of standard deviation from successive differences. It can be shown that

$$\sigma(\delta) = \sqrt{\frac{\overline{\delta^2}}{2}}$$

This statistic is a powerful tool for measuring the variation due to a system of chance causes, even though variation due to other causes may be present in the data. It is also of service in detecting the presence of assignable cause of variation even when such causes may not be detected by dividing the data into supposedly rational subgroups by the control chart technique previously outlined. Its properties can be brought into forms by a few comments on its origin which are well merited, since this statistic appears to have little background of use outside of military circles.

The probable error of artillery matériel is determined by firing a number of shots from the matériel in question, measuring the distance to the points of impact, and computing the estimated probable error. Dr. A. A. Bennett, who was on duty at the Ordnance Office and later at the Proving Ground during World War I, realized that the probable error estimated from

$$\text{P.E.} = 0.6745 \sqrt{\frac{\sum^{n}(\bar{X} - X_i)^2}{n - 1}}$$

was sometimes fictitiously large owing to weather conditions surrounding the experiment; e.g. an increasing tail wind would introduce a shifting mean and, hence, increase the second moment of points about the mean of all points. He therefore introduced the method of estimating P.E. from $\sigma(\delta)$,

$$\text{P.E.} = 0.6745\ \sigma(\delta),$$

in order to minimize external effects.

Recently, Mr. R. H. Kent of the Ballistic Research Laboratory and Professor J. von Neumann of the Institute for Advanced Study, in an unpublished paper, rigorously deduced the formula for the average squared standard deviation estimated from successive differences, $\overline{\sigma^2(\delta)}$. Bennett's result was right [8] although his procedure

[8] Bennett also deduced P.E. = 0.6 times the mean successive difference. This relationship is also shown without proof in Chapter XI, C. Cranz and K. Becker, *Hand Book of Ballistics*, H. M. S. Stationery Office, London, 1921. The efficiency of the statistic as shown in Cranz is wrong. Helmert's work, to which he refers, treats of the differences of all the pairs.

was in error. Subsequently, the author deduced a proof [9] which requires only elementary methods.

The exact distribution of $\sigma^2(\delta)$ is not known, nor are all its moments known. However, von Neumann and Kent have shown that the variance of $\sigma^2(\delta)$ is

$$\sqrt{\frac{3n-4}{(n-1)^2}\,\sigma'^4}.$$

The variance of the estimate of variance, i.e., of $[(n-1)/n]\,\sigma'^2$, is

$$\sqrt{\frac{2(n-1)}{n^2}\,\sigma'^4}.$$

The ratio of these variances, relative to the squares of the expected values of the estimates themselves, is a measure of their relative efficiency.

$$\text{Eff.} = \frac{\dfrac{2(n-1)\sigma'^4}{n^2} : \dfrac{(n-1)^2}{n^2}\,\sigma'^4}{\dfrac{3n-4}{(n-1)^2}\,\sigma'^4 \div \sigma'^4}$$

This reduces to:

$$\text{Eff.} = \frac{2n-2}{3n-4} = \frac{2}{3}\left[1 + \frac{1}{3n-4}\right].$$

It is thus obvious that, for large n, the efficiency of the estimate from the mean square successive differences, relative to that from

[9] Given n observations $X_1, X_2, \cdots, X_n$, n being even. Divide the data into k_a ($k_a = n/2$) subgroups of 2, in the order of occurrence,

$$\sigma_a^2 = \frac{1}{k_a}\sum_{j=1}^{k_a}\sum_{i=1}^{2}\frac{(\bar{X}_j - X_{ij})^2}{2-1}.$$

This is an unbiased estimate of $(\sigma')^2$. Divide the data into k_b ($k_b = [n-2]/2$) subgroups of 2, in the order of occurrence, omitting the first and last observations.

$$\sigma_b^2 = \frac{1}{k_b}\sum_{j=1}^{k_b}\sum_{i=1}^{2}\frac{(\bar{X}_j - X_{ij})^2}{2-1}.$$

This is also unbiased estimate of $(\sigma')^2$. The weighted mean of σ_a^2 and σ_b^2 is an unbiased estimate of $(\sigma')^2$.

$$\frac{k_a\sigma_a^2 + k_b\sigma_b^2}{k_a + k_b} = \text{est. } (\sigma')^2.$$

By substitution of values and a little rearrangement this reduces to $\sigma^2(\delta)$. For n odd, the proof is similar.

variance, is approximately 67%, and that as n decreases the former estimate becomes more efficient. For $n = 2$, the efficiencies are, of course, equal. It may be concluded, therefore, that, for small samples, the standard deviation by successive differences does not appear to lose a great deal in efficiency. This observation, of course, is subject to limitations which might result from the nature of the distribution. However, it has been confirmed by rather extensive empirical checks that, with samples of the order of size 20, $\sigma(\delta)$ appears to have about the same precision as σ computed from approximately $\frac{2}{3}$ the sample size. The efficiency of $\sigma(\delta)$ is appreciably superior to standard deviation from k subgroups of 2, as can be seen from comparing the above ratio for samples of 20 with the standard deviation of 10 samples of 2 as shown by Fig. 12·1.

In addition to its use for measuring the variation of a chance cause system (like the distances of shots from a gun) as if it were independent of an external cause system (like the shifting mean), the standard deviation by successive differences can be used as a check on the rationality of subgroups.

Given k subgroups of m, $n = km$. It is well known that, if the standard deviation computed over the whole n observations is significantly greater than the average standard deviation of the subgroups divided by c_2, one may doubt statistical control even though no points on the control chart are outside of limits. That is,

$$\sigma_n > \frac{1}{c_2}\,\bar{\sigma}_m$$

indicates[10] lack of control ($>$ meaning significantly greater than). $\sigma(\delta)$ having been computed for each of k subgroups in like manner,

$$\frac{1}{c_2}\,\bar{\sigma}_m > \sqrt{\frac{1}{k}\sum_{j=1}^{k}\sigma_j^2\,(\delta)}$$

indicates[10] a changing cause system within one or more of the subgroups; i.e., the subgroups are not rational in the sense that the cause system is constant within the subgroup. This may lead to the discovery of rational subgroups (through the method of trial and error) which will lead to the discovery of the assignable cause for variation.

[10] A beautiful measure of this character can be performed, using variance and the z-test as a measure of significance, see Chapter 23 of Yule and Kendall, *op. cit.*

Correlation, general. To the student of general statistics it may seem unusual to relegate correlation to miscellaneous techniques. It is first in appeal to the imagination, it is widely used in so many fields, and it is the subject matter of approximately twenty-five per-cent of the literature. The reasons for this apparent slight are three-fold. First, occasions for its use in engineering work are relatively infrequent, and, even when they arise, unaided engineering judgment is likely to be more nearly self-sufficient than in the cases previously discussed. Second, correlation, without both a competent knowledge of statistical theory and a mastery of the technical features of the scientific or engineering field to which the data pertain, may lead to erroneous conclusions. It is important that statistics lead to valid conclusions and not be used through design or ignorance as a support for unwarranted assumption. Third, a comprehensive procedure for the occasional user of statistical methods cannot be offered in a clear-cut and ready form.

The meaning of correlation. Correlation is an expression of relationship. This idea is brought more clearly into focus if one first considers a simple functional relationship such as $y = f(X)$. In this case, for every X there is a corresponding Y. For example, consider the formula:

$$L = 2\pi r.$$

For every r there exists one and only one L.

Now suppose that, instead of an r giving rise to one and only one L, the selection of an r gives rise to the selection of an L from a frequency distribution in which, let us say, the *mean* L is a function of r. Then the values are statistically related, and two values which are statistically related are said to be correlated.

Correlation implies commonness of causation with respect to two or more phenomena. This does not necessarily mean a relationship of cause and effect, nor is it even necessary to know the reason for or technological explanation of the commonness of causation, if the data are extensive enough. For example, in certain types of metals there is a high degree of correlation between hardness and tensile strength. This relationship is so reliable that in many types of commercial work hardness is tested rather than tensile strength, especially in instances where the latter test would be destructive as testing the strength of automobile axles. Of course, there is seldom a one-to-one correspondence between phenomenon A and phenomenon B. If there were,

the relationship would be functional and not statistical. That is to say, a hardness reading of 70 (Rockwell) may on the average correspond to a tensile strength of 32,000 pounds per square inch but may be expected to be as high as 40,800 or as low as 23,200 pounds per square inch. A measure of this scatter will be given shortly. Therefore, in using the indirect but correlated test, allowance must be made for the lack of perfect correlation between the two phenomena. Perfect correlation is indicated by a correlation coefficient of 1.0; no relationship between phenomena is indicated by a correlationship coefficient of 0; and a perfect negative relationship (in the sense that one variable increases as the other decreases) is indicated by a correlation coefficient of -1.0.

However, the observation of a moderate degree of correlation does not necessarily mean that commonness of causation exists. The observation may be due to chance. Likewise, the observance of zero correlation does not necessarily mean that no correlation exists.

Nature of the correlation coefficient. It has been shown for various statistics that, if the true value of the statistic is θ', the observed value of θ in samples of size n may be greater or less than θ', and on the average may or may not approach θ' as a limit. For example, the observed values of σ in small samples from a normal universe do not approach σ' on the average but require the correction factor $1/c_2$. A similar condition holds for the correlation coefficient, r. Shewhart [11] has shown that, in the actual drawing of 86 samples of 25 from a normal universe of $r' = 0$, the observed value of r varied from -0.60 to $+0.30$. However, in 86 samples of 25 from a normal universe in which $r' = 0.98$, r varied only from 0.95 to 0.992. The lesser dispersion of r as r' approaches unity is obvious from the approximate formula for the standard deviation of r,

$$\sigma_r = \frac{1 - r'^2}{(n-1)^{\frac{1}{2}}}.$$

The formula also indicates reduced dispersion as n increases. R. A. Fisher [12] published the distribution of r in 1915.

Calculation of the correlation coefficient. Given n pairs of values X_1Y_1, $X_2Y_2 \cdots X_nY_n$,

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sigma_X \sigma_Y},$$

[11] *Loc. cit.*

[12] *Biometrika*, Vol. X, 1915, page 507, et seq.

where $\overline{XY}$ = the average of the products of the pairs.
$\overline{Y}\overline{X}$ = the average of the X's times the average of the Y's.
σ_X = the standard deviation [13] of the X's.
σ_Y = the standard deviation [13] of the Y's.

Consider the following example. A certain component mechanism of a fuze was subject to test by measuring the number of pounds (force) required to cause the mechanism to function. Owing to difficulties associated with this type of test, it was desired to substitute in lieu thereof a simple test of resistance of the mechanism to a falling weight. The substitution required a knowledge of the statistical relationship between falling weight and pounds of force as applied to this particular mechanism. The data of Table 12·3 are a part of the test data [14] for this purpose.

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2} = \sqrt{\frac{2266.70}{20} - (10.55)^2}$$

$$= \sqrt{113.34 - 111.30} = 1.42.$$

$$\sigma_Y = \sqrt{\frac{133616.22}{20} - (80.55)^2} = 13.88 \quad \text{(see note 15).}$$

$$r = \frac{\overline{XY} - \overline{X}\overline{Y}}{\sigma_X \sigma_Y} = \frac{868.32 - (10.55)(80.55)}{(1.42)(13.88)} = 0.937.$$

By using this value of r as an estimate of true r', the standard deviation of r can be calculated by the approximate formula

$$\frac{1 - r'^2}{(n-1)^{\frac{1}{2}}} = \frac{1 - 0.937^2}{(19)^{\frac{1}{2}}} = 0.028.$$

[13] In this case the correction factor c_2 should be omitted in the calculation of σ_X and σ_Y, because c_2 should be applied only in connection with linear functions of the standard deviation, and generally only in connection with control chart technique. Another formula for r which avoids this issue is

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 \Sigma(Y - \overline{Y})^2}}.$$

[14] The data have been altered but not in a way to prejudice the illustration.

[15] $$\sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}} = \sqrt{\frac{\Sigma X^2}{n} - \frac{2\overline{X}\Sigma X}{n} + \overline{X}^2} = \sqrt{\frac{\Sigma X^2}{n} - \overline{X}^2}.$$

It should be observed that in the notation employed the capital letters refer to observations whereas the small letters refer to the difference between the observations and their mean. Thus $\overline{X}$ refers to the mean of the X's but $\bar{x}$ to the mean of $(\overline{X} - X_i)$.

If one recalls that positive correlation varies between 0 and 1, this appears to be a rather high degree of correlation. It is difficult to say, however, just what this means in a clear and concise way. At least a part of this difficulty is associated with the nature of r itself. The correlation coefficient is of considerable value inasmuch as it enables

TABLE 12·3

COMPUTATION OF CORRELATION FOR INCHES DROP AND POUNDS FORCE

Inches drop of $\frac{1}{4}$-lb. ball X	Force in Y	X^2	Y^2	XY
7.13	48.00	50.77	2304.00	342.00
7.87	59.25	62.02	3510.56	466.59
9.19	60.75	84.41	3690.56	558.14
10.50	68.25	110.25	4658.06	716.63
9.19	71.25	84.41	5076.56	654.61
10.13	73.50	102.52	5402.25	744.19
10.69	75.75	114.22	5738.06	809.58
10.12	78.75	102.52	6201.56	797.34
9.94	79.50	98.75	6320.25	790.03
10.31	81.75	106.35	6683.06	843.05
10.31	83.25	106.35	6930.56	858.52
11.44	84.00	130.82	7056.00	960.75
11.06	86.25	122.38	7439.06	954.14
10.88	87.75	118.27	7700.06	954.28
10.31	88.50	106.35	7832.25	912.66
12.00	90.00	144.00	8100.00	1080.00
12.00	93.00	144.00	8649.00	1116.00
12.56	97.50	157.82	9506.25	1224.84
12.37	99.75	153.14	9950.06	1234.41
12.94	104.25	167.38	10868.06	1348.73
Total 210.94	1611.00	2266.70	133616.22	17366.49
Aver. 10.55	80.55	113.34	6680.81	868.32

one to express the degrees of relationship between two series of data in general numerical terms which are independent of the units of the original data. It does not, however, supply any means of passing over from one series of data to corresponding values of the other series, nor does it give any definite knowledge of the loss in precision that should be expected if such a transfer were attempted by some known process. Sometimes the mere expression of the degree of

relationship r is surely an advantage. However, from the engineering point of view it appears much more important to know the best way to interpret one series in terms of the other, i.e., pass over from one

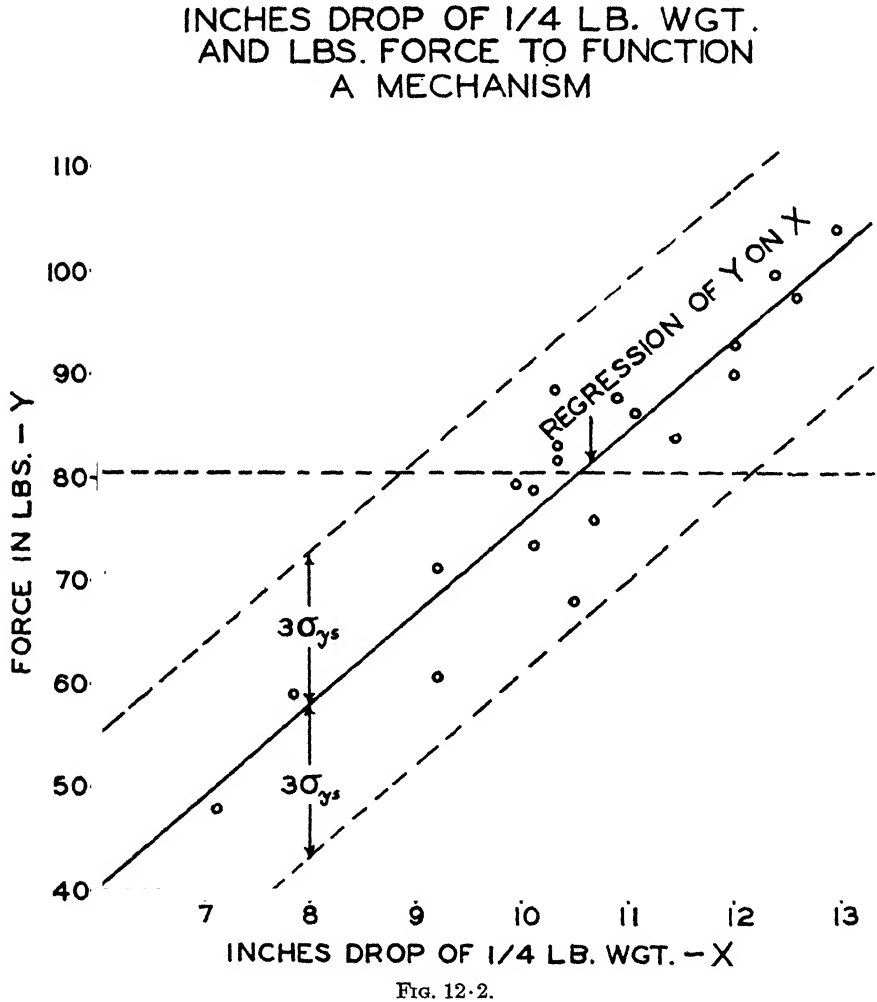


FIG. 12·2.

series of data to the other, and to have a measure of the precision of this type of interpretation. It will be seen that this is a relatively simple procedure.

Estimating one series of data in terms of a correlated series. Figure 12·2 shows a plot of the data of Table 12·3. A very rough

engineering way of interpreting one measure in terms of the other would consist of drawing a line by eye through the scatter of points so as to appear to approximate the points as closely as possible. A more elegant and frequently used procedure would consist of fitting the points with a least squares line so as to make the sum of the squares of the distances of the points from the line (measured perpendicular to the line) a minimum. This is known as the line of best fit. However, if one considers the fact that X is the independent variable which is going to be directly measured and hence has no error about it (except errors of observation) and that Y, the dependent variable, is statistically related to X, then it is obvious that it is better to minimize the Y errors (the X errors already being small). Hence, it is far better to use, as an estimating line, a line that makes the sum of the squares of the vertical distances of the points from the line a minimum. This is known as the regression line of Y on X. The equation of the line (assuming that the relationship is linear) is obviously of the form

$$Y = a + bX,$$

and the constants a and b can be evaluated by solving the equations:

$$\Sigma Y = na + b\Sigma X,$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2.$$

After a systematic tabulation of X, Y, X^2, Y^2, and XY, the solution of these equations is easy. In the present case the indicated summations can be taken from Table 12·3.

$$1611.00 = 20a + 210.94b,$$

$$17{,}366.49 = 210.94a + 2266.70b.$$

Solving, one gets

$$Y = -12.74 + 8.85X.$$

Now, if the statistical relationship between X and Y is such that the standard deviation of the points on any ordinate erected on Fig. 12·2 is equal to the standard deviation of the points on any other like ordinate [16] (measurements being made from the regression

[16] This is the usual assumption, and the distribution is said to be homoscedastic. In this example, paucity of data and method of measurement cause some ordinates to have 0 or 1 point.

line of Y on X), then the standard deviation of the Y values about the regression line, σ_{ys}, is

$$\sigma_{ys} = \sigma_y\sqrt{1 - r^2}.$$

Thus, in the present case $\sigma_{ys} = 13.88\sqrt{1 - 0.937^2} = 4.86$. By this procedure one can now put in limits at $\pm 3\sigma_{ys}$ on Fig. 12·2 which are similar to control chart limits and within which practically all observed Y values should fall. The standard deviation σ_{ys} gives a measure of the variation or error that may result from estimating Y from X. Thus, by the procedure just outlined, one has a method of estimating Y from X and a measure of the range of variations in such estimates.[17]

In like manner one could, of course, regard Y as the independent variable and X as the dependent variable, obtain the best line for estimating X from Y which is the regression line of X on Y, and derive σ_{xs}. Then the line would be different, and the equations are:

$$X = a + bY,$$
$$\Sigma X = na + b\Sigma Y,$$
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2,$$

$$\sigma_{xs} = \sigma_x$$

The two regression lines coincide if, and only if, $r = 1$. The best representative line (minimizing the sum of the squares of the deviations perpendicular to the line) lies between the two regression lines, and all three lines pass through the point, $\overline{X}$, $\overline{Y}$.

Comments on correlation techniques. If the pairs of observations were 100 or more, the procedure just outlined would become quite laborious. Under these circumstances much labor can be saved by grouping data into cells as explained in various elementary textbooks on statistics. It is suggested that the occasional user of statistical methods who is not familiar with such procedure avail himself of the services of someone who is, rather than incur the labor of going into such a matter for the sake of a few isolated applications.

[17] Chapter XXII of *Applied General Statistics*, Croxton and Cowden, gives an illuminating explanation of these variations in terms of variance (the squared standard deviation). σ_y^2 is regarded as the total variance of the Y values about their mean. Of this total variation, the amount $r^2\sigma_y^2$ is explained by the correlation coefficient, hence is accounted for. The remainder of the total variance, σ_{ys}^2, is regarded as "unexplained variance" and is the amount we must allow for in making estimates.

$$\sigma_y^2 = \sigma_{ys}^2 + r^2\sigma_y^2.$$

In the example cited it was assumed that the statistical relationship was linear. This assumption appears to be a close approximation in the practical working range, but it is obviously not true in principle as can be seen by contrasting the engineering certainty, $X = 0$, $Y = 0$, with the regression equation, $Y = -12.75 + 8.85X$. That is to say, at least the lower left-hand portion of the line must become curvilinear and concave upward. Whereas in many cases linear correlation is either correct or a sufficiently close approximation in the working range, instances arise where a correlation obviously exists which is non-linear in the working range.

Whereas it is possible to fit data with curves of any order, the treatment of non-linear regression is distinctly a task for the statistician or mathematician rather than the occasional user of statistical methods. However, a large proportion of these apparently difficult problems will yield to the elementary methods just outlined, if a little ingenuity is employed. It is suggested that at least the following simple steps be tried:

(*a*) Plot the data on semi-logarithmic paper with X on the logarithmic scale.
(*b*) Try Y on the logarithmic scale.
(*c*) Try the data on paper with both axes to logarithmic scale.

By proper choice of scales many curves can be changed to approximately straight lines. Obviously, the above procedure takes care of curves of the forms:

$$Y = AB^X \text{ and } X^nY = A$$

Note that the correlation is between one variable and the logarithm of the other and between the logarithms of the variables, respectively, and calculations must be performed with these respective values, not the original data.[18]

By way of illustration consider two series of measurements [19] of the blast effect of an explosion as shown in Table 12·4. The measurements were made at various distances from the explosion by two types of instruments: instrument type A and instrument

[18] Linearity can sometimes be improved by multiplying one variable by a constant, and handling the new variables X and CY; or by substituting a new variable, Z, for XY in a relationship of the form $XY = A + BX$. Tests of linearity are beyond the scope of this book.

[19] These data have been altered but not in such a way as to prejudice the illustration.

type N. From the engineering principles of the instruments it appeared that both should measure the same thing, viz., the integral of pressure times differential time, $\int Pdt$, and hence that the measurements should be correlated.

TABLE 12·4

MEASUREMENT OF UNITS OF BLAST EFFECT BY TWO TYPES OF INSTRUMENTS

Units Inst. Type A	Units Inst. Type N	Units Inst. Type A	Units Inst. Type N
30.0	7	8.5	3
25.0	7	8.0	3
17.5	5	7.7	2
19.3	5	7.5	2
14.5	5	6.5	1
11.5	4	5.0	0
10.8	4		

A plot of the readings, Fig. 12·3A, shows that the correlation, if any, is distinctly curvilinear. However, when plotted on semi-logarithmic paper, the relationship appears to be clearly linear (see Fig. 12·3B).

In this case $\overline{X}^2 = 2.2476$, $\overline{X^2} = 1.1746$.

$$\overline{Y}^2 = 13.6331,\ \overline{Y^2} = 17.8462.$$

$$\overline{XY} = 4.3700,\ \overline{X}\,\overline{Y} = 3.9109.$$

$$\sigma_x^2 = \overline{X^2} - \overline{X}^2 = 0.0527,\ \sigma_x = 0.2295.$$

$$\sigma_y^2 = \overline{Y^2} - \overline{Y}^2 = 4.2131,\ \sigma_y = 2.0526.$$

$$r = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_x\,\sigma_y} = 0.9745.$$

The correlation between the reading of instrument type N and the logarithm of the reading of instrument type A therefore appears to be very high. The equation of the regression line of the reading of instrument type A on the logarithm of the reading of instrument type N, i.e., the equation for changing observations of N to corresponding observations of A, is

$$A = -5.55 - 8.72 \log N.$$

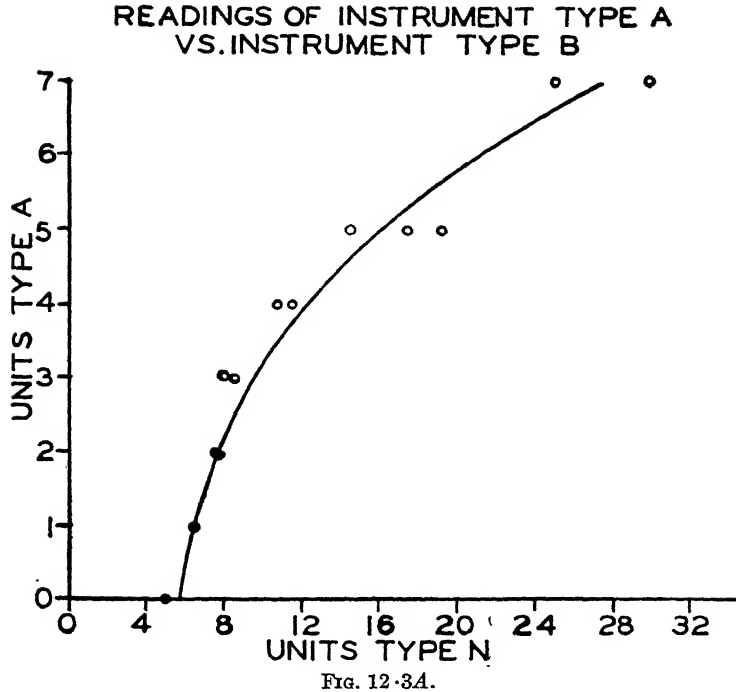


FIG. 12·3A.

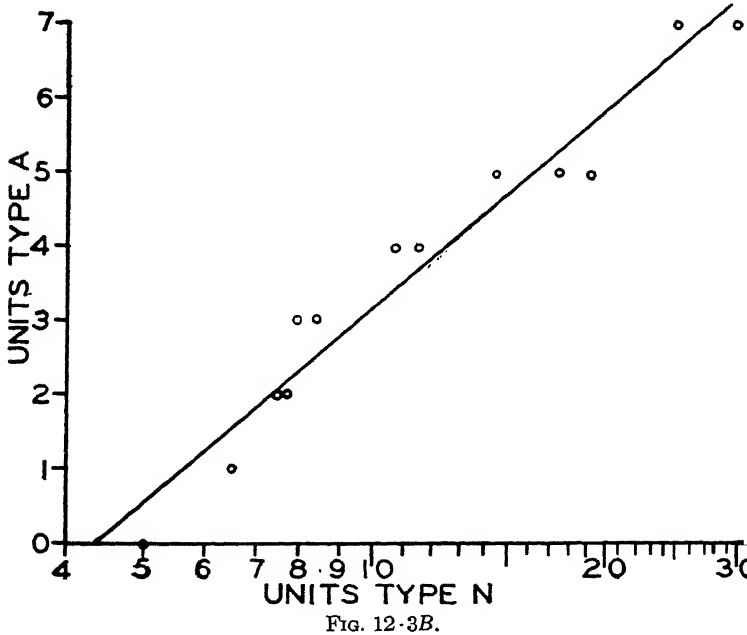


FIG. 12·3B.

Hence, the correlation between the observed value of one variable and the logarithm of the other is no detriment to interpreting one variate in terms of the other.

TABLE 12·5

PRECISION OF ANALYSIS OF SMOKELESS POWDER DETERMINED BY DUPLICATE SAMPLES[20]

Sample No.	Result of sample A, %	Analysis sample B, %	Absolute difference		
			D	$\bar{D} - D_i$	$(D - D_i)^2$
1	4.64	4.86	0.22	0.06	0.0036
2	4.71	4.73	0.02	0.14	0.0196
3	4.71	5.03	0.32	0.16	0.0256
4	4.93	4.73	0.20	0.04	0.0016
5	4.72	4.78	0.06	0.10	0.0100
6	4.98	4.91	0.07	0.09	0.0081
7	4.90	4.68	0.22	0.06	0.0036
8	4.81	4.73	0.08	0.08	0.0064
9	5.03	4.35	0.48	0.32	0.1024
10	4.68	4.81	0.13	0.03	0.0009
11	4.80	4.84	0.04	0.12	0.0144
12	4.95	4.92	0.03	0.13	0.0169
13	4.82	4.62	0.20	0.04	0.0016
14	5.00	5.04	0.04	0.12	0.0144
15	4.80	4.86	0.06	0.10	0.0100
16	4.83	4.70	0.13	0.03	0.0009
17	4.90	4.60	0.30	0.14	0.0196
18	4.66	4.76	0.10	0.06	0.0036
19	4.73	4.87	0.14	0.02	0.0004
20	4.76	4.66	0.10	0.06	0.0036
21	4.48	4.74	0.26	0.10	0.0100
22	4.80	4.72	0.08	0.08	0.0064
23	5.01	4.65	0.36	0.20	0.0400
			Total 3.64		0.3236
			Aver. 0.16		0.01407

It should be emphasized that the significance of any measure of relationship is highly dependent upon the form of the relationship; and, in contrast to this observation, it is well to bear in mind the fact that the simple statistics $\overline{X}$ and σ convey so much information which

[20] Original data have been altered but in such a way as not to be prejudicial to the illustration.

is independent of the form of the distribution and is made useful through the Tchebycheff theorem.

Measurement of precision of observation of a variable. In engineering work it is often desirable to know the precision of observation of some phenomenon. The statistical method offers a ready means of measuring the precision of an observing process even though the true objective measure of the phenomenon is never known. For example, suppose that the acceptance of a product requires that the chemical analysis of a sample be between certain limits. If the chemical determination is imprecise, the product may be rejected because of errors of observation. The method of checking the precision of observation can be easily shown by an example.

In the manufacture of smokeless powder, acceptance of lots of powder is dependent upon the maintenance of the amount of one of the chemical ingredients at less than an assigned percentage. It was suspected that lots were being rejected because of lack of precision in chemical analysis of samples. The chemical laboratory insisted that its results were precise beyond question. In order to test this stand, duplicate samples were taken from each powder lot, without permitting the laboratory to know which samples were duplicates. The duplicate samples were not two random samples from the same lot, as such might differ considerably from each other, but halves of a single sample obtained by the process of mixing and quartering. Hence the duplicates were in fact duplicates in the sense of being almost precisely the same. The results are shown in columns 2 and 3 of Table 12·5.

Since the true but unknown percentage is the same in sample A and sample B, their difference is due to accidental errors of observation which occur in the two independent tests of the duplicate samples. That is, the variable shown by the fourth column of Table 12·5, viz. D, is the algebraic sum of two independent variables, viz., the first analysis and the second analysis. The two independent variables, however, are equal, since each is the product of the same operational process.

$$\text{The std. dev. of the difference} = \sigma_D = \frac{1}{c_2}\sqrt{\frac{\Sigma(D - D_i)^2}{n}}$$

$$\frac{1}{0.9555}\sqrt{0.01407} = 0.1236.$$

The std. dev. of the error of observation $= \sigma_a$.

It can easily be shown that the standard deviation of the sum of two independent variables [21] is equal to the square root of the sum of the squares of the standard deviations of the respective variables; i.e.,

$$\sigma_D = \sqrt{\sigma_{aA}^2 + \sigma_{aB}^2}.$$

Since $\sigma_{aA} = \sigma_{aB} = \sigma_a$, $\sigma_D = \sigma_a\sqrt{2}$, therefore

$$\sigma_a = \frac{\sigma_D}{\sqrt{2}} = \frac{0.1236}{1.414} = 0.0874.$$

Since the standard deviation of the observation is 0.09%, it is easily seen that reported results might well be in error by as much as ±0.27%. After this demonstration the laboratory was willing to run check tests in cases which were close to the limiting value.

It should be noted that this procedure measures the precision (reproducibility) of the test, not its accuracy (closeness to the true value). For example, suppose that the tests consisted of measuring the loss of weight of the sample after desiccation and that desiccation was always incomplete. There could well be a mean constant error in results, and no statistical means would detect the error. Precision can generally be measured by statistical means; measurement of accuracy is often more difficult. A measure of accuracy generally involves some engineering technique. For example, in the present case it might be accomplished by making a synthetic sample of known percentage of the ingredient in question, then subjecting the known sample to the analysis, and finally comparing results of analysis with the known percentage. In general, a test of accuracy requires an artificial set-up in which the answer is known before the test is made.

Further use of the sum of two independent variables. Consider a universe composed of k universes which are identical except for the values of their respective averages. Let the universes be controlled about the averages, $X_1, X_2, \cdots, X_k$; and each with standard deviation σ_a, where σ_a can be considered the variation due to accidental causes. Let the $\bar{X}$'s of the respective universes be controlled about a mean

[21] The standard deviation of a linear function of any number of variables is

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 \cdots + \sigma_n^2 + 2r_{1,2}\sigma_1\sigma_2 + \cdots 2r_{n-1,n}\sigma_{n-1}\sigma_n},$$

where r is the correlation coefficient. (For independent variables, $r = 0$.) See Yule and Kendall, Chapter 16.

value $\bar{\bar{X}}$ with standard deviation σ_s, where σ_s is used instead of $\sigma_{\bar{X}}$ to promote thinking of the variation as due to systematic causes, i.e., a cause which is constant for a set of X's measured about a specific $\bar{X}$.

As an example of this type of compound variation, consider a product manufactured by k automatic machines all of which have been set by an operator or operators at an objective value of $\bar{\bar{X}}'$. Suppose further that the settings are randomly distributed about $\bar{\bar{X}}$ and that the variation in the product of each machine about its respective $\bar{X}$ is essentially the same and equal to σ_a. How can one predict the variation, σ_T, in the product resulting from the operation of all the machines on a basis of n samples from each machine at each setting?

The total variation is equal to the sum of two variations, i.e.,

$$\sigma_T = \sqrt{\sigma_s^2 + \sigma_a^2}.$$

Let X_{ij} be the ith observation in the jth subgroup. With respect to σ_a one has nk observations consisting of k subgroups of sample size n. One can, therefore estimate [22] σ_a^2 from $\frac{1}{k}\sum_{j=1}^{k} \sigma_j^2$, where

$$\sigma_i'^2 = \frac{\sum_{i=1}^{n} (\bar{X}_j - X_i)^2}{n - 1}$$

However, with respect to true $\bar{X}$'s, one has no observations whatsoever. Instead, one has only estimates each of which is based on n measurements of $\bar{X}_j$. Therefore $\sigma_{\bar{X}}^2$ (or rather σ_s^2) computed from the observed $\bar{X}$'s will be too large, as it contains not only the variation in $\bar{X}$ but also the variation due to the errors of measurement. Therefore

$$\sum_{j=1}^{k} \frac{(\bar{\bar{X}} - \bar{X}_j)^2}{k - 1} = \sigma_s^2 + \left(\frac{\sigma_a}{\sqrt{n}}\right)^2,$$

[22] It was shown by Gauss and later by R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, London, that the formula

$$\sigma'^2 = \frac{\sum_{i=1}^{n} (\bar{X} - X_i)^2}{n - 1}$$

yields an unbiased estimate of σ'^2. These remarks assume a normal universe.

where $(\sigma_a/\sqrt{n})^2$ is, of course, the squared standard deviation of an average of n, and the total variation measured is the sum of the variations due to the systematic causes plus the variation due to the accidental errors in an average of n.

Solving for σ_s^2, one gets

$$\sigma_s^2 = \sum_{j=1}^{k} \frac{(\overline{\overline{X}} - \overline{X}_j)^2}{k-1} \qquad \frac{\sigma}{n}$$

It follows, therefore, that

$$\sigma_T^2 = \sigma_s^2 + \sigma_a^2 = \left(\sum_{j=1}^{k} \frac{(\overline{\overline{X}} - \overline{X}_j)^2}{k-1} - \frac{\sigma_a^2}{n}\right) + \sigma_a^2$$

$$= \sum_{j=1}^{k} \frac{(\overline{\overline{X}} - \overline{X}_j)^2}{k-1} + \frac{n-1}{n} \cdot \sigma_a^2.$$

Now note that in this case σ_a^2 is used and that one is making no practical use of σ_a. The formula

$$\sigma = \frac{1}{c_2} \sqrt{\sum_{i=1}^{n} \frac{(\bar{x} - x_i)^2}{n}}$$

gives an unbiased estimate of the standard deviation, σ. It does not give an unbiased estimate of σ^2. It can be easily shown that the formula

$$\sigma^2 = \sum_{i=1}^{n} \frac{(\bar{x} - x_i)^2}{n-1}$$

gives an unbiased estimate of σ^2. Since the mean of the squares of two or more numbers is not necessarily equal to the square of their mean, it follows that in the above case it was better to use the formula for the unbiased estimates of the squared standard deviations.

Observing the notation previously indicated, and for brevity allowing $d_i = \overline{X}_j - X_i$ and $D_j = \overline{\overline{X}} - \overline{X}_j$, one can then write

$$\sigma_T = \sqrt{\sum_{j=1}^{k} \frac{D_j^2}{k-1} + \frac{n-1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{d_{ij}^2}{k(n-1)}}$$

$$= \sqrt{\sum_{j=1}^{k} \frac{D_j^2}{k-1} + \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{d_{ij}^2}{kn}}.$$

If kn in the last expression is changed to $(kn - 1)$, this is the result that would have been obtained if est. σ_T' had merely been computed from the whole of the observations in the usual way [23] for computing an unbiased estimate of σ_T^2. This is as one would expect, and the advantage of the illustrated technique lies not in its peculiar efficacy in computing σ_T, but in the ability to compute σ_T from its components. All control technique would ordinarily be centered on the components, since analysis by each component is the best way to detect and eliminate the assignable causes for variation. One can then get σ_T, if one desires, without an additional, extended computation.

This technique is of especial value in the analysis of bombing data, where the two types of variation just pointed out comprise the most important way in which the underlying principles of bombing differ from those of gunfire.

If the n's are not equal, and if one wishes to introduce the question of weighting factors, complications arise which are beyond the scope of this book. It may be observed, however, that if the n's are not equal the introduction of weighting factors is not likely to do much good in the practical work, however much one may wish for meticulous precision.

[23] Because $J_D = J_0 + D^2$, where J_D = the second moment about a point D distance from the mean and J_0 is the second moment about the mean.

It is my feeling that the chief duty of a statistician is to interpret data in such a way that they convey knowledge for the purposes of prediction; another is to collect the data in such a way that they provide the maximum information; and still a third function is to help bring about some changes in the source of the data. There is no use in taking data, if you do not intend to do something about the sources of the data.

—Letter from Dr. W. Edwards Deming to the Chairman of the War Preparedness Committee of the Institute of Mathematical Statistics, October, 1940.

APPENDIX A

ESTIMATING THE FRACTION OF A LOT WHICH POSSESSES A PARTICULAR QUALITY CHARACTERISTIC FROM SAMPLES

General discussion. The one certain way to determine the fraction effective of a lot of articles is to test each article thereof under appropriate conditions. However, inspection and testing are always expensive, and in cases where the test is destructive, 100% inspection is, of course, impossible. It therefore follows that in most practical work lot quality is estimated from sample quality, and one must accept an assurance somewhat less than certainty that the estimate is correct.

Chapter I threw some light on this sampling problem and showed that small samples from a moderately defective lot of articles are better than the lot more frequently than they are poorer than the lot. Chapter II gave a method of allowing for the bias of the sample and of estimating the precision of the estimate, viz., estimating the probable limits within which the lot fraction effective can be expected to lie. Neither chapter dealt with the fundamental principles underlying the problem. The problem can be divided into three parts: first, the design of the sampling scheme so that known laws of probability apply; second, either knowledge or a hypothesis regarding the distribution of the kind of lot or lots sampled; and third, the mechanism for readily handling numerical computations. This appendix proposes to discuss these fundamental principles.

Randomness of Sample. Differences of opinion with regard to randomness and its relation to analysis are often attributable to incomplete understanding of a few basic concepts and their relation to analytical techniques. A sample has been selected in a random manner when all articles of the lot were equally likely to be selected during the sampling process. For brevity, such a sample is called (with doubtful propriety) a random sample. However, it is important to note that the *randomness pertains to the operation of selecting the sample,* and not to the sample itself; and that any sample whatever can be arrived at either by a random or non-random operation.

Homogeneity of lot bears an important relation to randomness of

sampling. The word homogeneous, as applied to a lot of articles which is actually heterogeneous in that the articles differ one from another, implies that these articles are throughly mixed. But if they are thoroughly mixed, then every article has an equal chance of being in any part of the lot; and, consequently, any portion of a homogeneous lot is, per se, the equivalent of a random sample. Therefore, it may be noted that the random operation of selecting the sample is tantamount to mixing the articles in the abstract. Furthermore, *if the sampling is random, it is not necessary that the lot be homogeneous; and if the lot is homogeneous, it is not necessary that the sampling be random.*

Footnote 3 of Chapter II should now be clear. Although the specific definition of random sample is violated, if the member lots are the same, then the grand lot is also the same as any member lot. Hence, the random sampling of the member lots is nothing more or less than the successive random sampling of the grand lot. But if the member lots are different, then the above equivalence does not hold, and one has simply an aggregate sample consisting of random selections from different strata of the grand lot.

The randomness-homogeneity relationship is the mainspring of the powerful control-chart technique. The small samples from the rational subgroups certainly are selected in a non-random manner with respect to the whole lot. The control chart is essentially a test of whether the samples appear to behave as random samples from a universe. If the non-randomly selected samples appear to behave as random samples, it follows that the universe appears to be homogeneous.

The assumption of randomness of sample is an inevitable *sine qua non* in a large part of the calculation of basic relations of sampling phenomena. It is essential that the engineering principles involved in lots of articles be carefully studied; and when practicable, that the sampling procedure be logically planned so as to take every reasonable precaution to insure randomness.

However, the absence of formal assurance of randomness does not necessarily preclude statistical aid. The engineer may insist that the statistical requirements are without important effect in his work; and reject physical mixing as impossible, mixing in the abstract as impractical, and sufficient samples for a statistical test of randomness as financial improvidence. His stand as he presents it may appear to be on intuitive grounds only, but actually there are three cogent

arguments in his behalf. The practical aspects of the problem are worthy of consideration.

First, since randomness of sample inheres in the operation of selecting the sample the most intimate evidence of randomness is the judgment of the experimenter. The man familiar with the technological aspects of the articles under consideration is in position to be a good judge — perhaps the only competent judge — of whether the selection was random or at least unbiased with respect to the quality characteristic under consideration.

Second, absence of homogeneity is not a valid a priori assumption. The practical man well knows, for example, that, under mass production, articles get quite a deal of mixing without either intent or design. They get mixed in the process of manufacture, mixed in the packing room, mixed as they are placed in the box car; and, if they are transshipped once or twice, they may be mixed almost as thoroughly as chips in a bowl. Ask the production man who has discovered that some defective articles have gone out and who has tried to recapture the articles.

Third, allowance must always be made for only partial fulfillment of hypotheses. In all sampling procedure, one or more hypotheses are implied. If one makes a test for a significant difference by a procedure that involves normal distribution, and if a significant difference is indicated, such a result merely means: (*a*) that the difference is significant; or (*b*) that the distribution is not normal; or (*c*) that the sampling was not random. In any event, the engineer must judge whether the conditions implied by the hypotheses are met, or approximately met, in practice. In the interest of precision and reproducibility, the judgments of the idealized conditions of hypotheses must be held to a minimum. Some of these judgments are very difficult to render. For example, the approximation of the distribution of the lot to normality could scarcely be judged without extensive statistical tests. This hypothesis should, therefore, be avoided when practicable. However, the assumption of randomness of sample can seldom be avoided. It appears that the engineer simply must include this consideration, when he selects his action criteria.

With respect to randomness of sample, one can neither throw discretion to the winds nor place undue burdens on practical execution. Even though the condition of randomness cannot be specifically checked, certainly this condition is no valid reason for abandoning otherwise satisfactory tests in favor of unaided and arbitrary decision.

Distribution of lots. Assuming that the sample satisfactorily approximates randomness, is one then in a position to predict lot quality from sample quality? It is suggested that the answer be deferred pending the consideration of a few examples. Chapter I showed how to calculate (*a*) the probability that a lot of certain fraction defective would produce any one of a number of kinds of samples, (*b*) the probability that a certain sample would be produced by any one of a number of kinds of lots. Let us now consider the probability that a certain sample came from a certain lot.

A posteriori probability. Consider a sample of 2 articles taken from a large lot of articles. Both articles prove to be good. The lot sampled was one of a stock of lots constituted as follows:

$$2 \text{ lots}, Q = 0.50;$$

$$6 \text{ lots}, Q = 0.25;$$

$$2 \text{ lots}, Q = 0.00.$$

One lot is as likely to have been sampled as another. What is the most probable quality of the lot? The procedure can best be shown in tabular form:

TABLE A1

CALCULATION OF A POSTERIORI PROBABILITY WHEN THE EXISTENCE PROBABILITY IS KNOWN

(1) Quality of lot P	(2) Probability that lot would produce the sample	(3) Probability that lot was sampled	(4) Product of the probabilities
0.50	$1/2 \times 1/2 = 1/4$	2/10	$1/4 \times 2/10 = 4/80$
0.75	$3/4 \times 3/4 = 9/16$	6/10	$9/16 \times 6/10 = 27/80$
1.00	$1 \times 1 = 1$	2/10	$1 \times 2/10 = 16/80$

Answer. It is most probable that the sample came from the lot which is $P = 0.75$, since it has 27/80 chances out of (4/80 + 27/80 + 16/80), or 57% against 9% and 34% for the other two kinds of lots. Hence, the most probable quality of the lot is $P = 0.75$, not $P = 1.00$, as one might suspect. It is suggested that the reader reflect on the logical relationship between this answer, the sample size, and the distribution of lots.

This is a very simple illustration of a posteriori probability,[1] i.e., the probability after the sample is taken. The great difficulty in a posteriori probability consists in knowing the distribution of lots, a factor which was given in this case. In general, this factor is not only unknown, but, if one took enough samples to determine the distribution with satisfactory approximation (assuming such sampling were possible), one would have enough sampling evidence regarding the respective lots without resort to any statistical methods.

TABLE A2

A Posteriori Probability, Existence Probability Unknown

(1) Lot in terms of P	(2) Probability that lot would produce the sample ($n = 10, c = 2$)	(3) Probability that lot was sampled	(4) Product of the probabilities
1.00	0.0000	1/100	0.000000
0.99	0.0042	1/100	0.000042
0.98	0.0153	1/100	0.000153
0.97	0.0318	1/100	0.000318
0.96	0.0520	1/100	0.000520
.	.	.	.
.	.	.	.
.	.	.	.
0.80	0.3020	1/100	0.003020
.	.	.	.
.	.	.	.
.	.	.	.
0.00	0.0000	1/100	0.000000

Let one consider then the sample ($n = 10$, $c = 2$) taken from a large lot, where the distribution of lots is not known. If nothing is known regarding the lot except the evidence of the sample, and if in the absence of knowledge one chooses to assume that one kind of lot is just as likely as another, then the probability that the sample came from any particular lot can be computed. This seems to be as reasonable an assumption [2] as one could make in a case of total ignorance

[1] An interesting discussion is given in *Calcul des probabilités*, 2 ième éd., H. Poincaré, Gauthier-Villars, Paris, 1912.

[2] It is not contended that ignorance is a logically valid basis for inference. It will be shown, however, that, within an experimentally limited range, it may serve a practically useful purpose.

regarding the stock, and it supplies that essential unknown regarding the prior conditions of sampling which is essential for proceeding from results to probable conditions, i.e., estimating the lot from the sample. The procedure on this basis for finding the probability that the sample ($n = 10$, $c = 2$) came from any lot is similar to that in the example just cited. In the former case the stock was composed of 10 lots: 2 with $P = 0.50$; 6 with $P = 0.75$; and 2 with $P = 1.0$. In the present instance (by way of convenient approximation) the stock may be considered as consisting of a large number of lots, an equal number of which are $P = 1.00$, $P = 0.99$, $P = 0.98 \cdots P = 0.01$, $P = 0.00$. A table similar to A1 can be constructed.

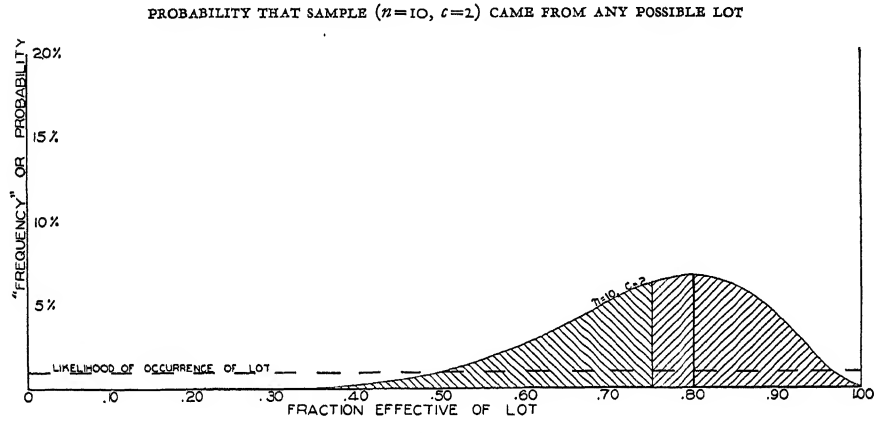


FIG. A1.

The probability that the sample ($n = 10$, $c = 2$) came from any lot is equal to its corresponding number in column (4) divided by the sum of the numbers in column (4). This is a simple illustration of Bayes' Theorem, discussions of which can be found in Shewhart,[3] Coolidge,[4] Keynes,[5] *Bell System Technical Journal*,[6] etc. The table, however, would be difficult to scan, and the results are much clearer in graphical form. Figure A1 is a curve showing the probability that the sample came from the lot which is 0.35, 0.36, 0.37 $\cdots$ 0.99

[3] *Loc. cit.*

[4] J. L. Coolidge, *An Introduction to Mathematical Probability*, Oxford Press, London, 1925.

[5] John Maynard Keynes, *A Treatise on Probability*, Macmillan and Co., London, 1921.

[6] E. C. Molina, "Bayes' Theorem: An Expository Presentation," *Bell System Technical Journal*, January, 1924.

fraction effective, each point on the curve being obtained in the manner just indicated.[7]

As may have been suspected, a lot of 0.80 fraction effective is the most probable lot, in the sense that of all possible lots it has the highest probability of having existed (under the basis of the hypothesis) and of having produced the sample. However, should one make a practice of making this type of decision, it is also obvious from Fig. A1 that one's estimates would in general be badly biased. (Note that the area to the right of lot 0.80 is much smaller than that to the left of this point.) If, however, one were to be faced with this decision once and only once, then the solution yielded by the modal or most likely value would appear logical. If one wished to make his estimates in such a manner that on the average the estimates would be closest to the true values, then, instead of selecting the modal value, he should select the average. This value would be the mean ordinate of the curve. It can easily be shown [8] that the average lot is the one whose fraction defective is [9]

$$Q = \frac{c + 1}{n + 2}.$$

In the present instance this lot would be $Q = (2 + 1)/(10 + 2) = 0.25$. However, in general, the action taken on estimates of this sort is predicated on arbitrary limits; i.e., one takes some specific action if the lot appears to be more defective than some predetermined value, and the action is the same whether the lot appears to be only slightly poorer than the criterion or a great deal poorer. Therefore, in order to make one's errors a minimum, one should select a value so that one's estimates will be too high just as frequently as too low. That is to say, one takes the average value with respect to frequency rather than with respect to magnitude. Obviously this is the median, or the value which divides the area under the curve into

[7] No discussion of a posteriori probability is offered in connection with continuous functions (sampling of variables), as other means of estimation are used. An interesting discussion is given in "The Frequency Distribution of the Unknown Mean of a Sampled Universe," E. C. Molina and R. I. Wilkinson, *Bell System Technical Journal*, Vol. 8, October, 1929.

[8] Coolidge, *loc. cit.*

[9] This is known as the law of succession. See Laplace, *Théorie analytique des probabilités*, Gauthier-Villars, Paris, 1886. It can be vigorously attacked as a general solution of this type of problem. See J. M. Keynes, *A Treatise on Probability*, Macmillan & Co., London.

two equal portions. In the present instance, this value is $P = 0.76$. Chart $0.5 = I_Q$ yields the solution of this problem sample sizes up to 500.

The effect of the a priori distribution of lots upon a posteriori probability. If one started to apply the procedure just outlined to a problem in one's professional field, it is quite likely that the assumption of a priori equal likelihood would appear to be an obvious fallacy. In most work one generally has at least some sort of general knowledge of the limits within which the kind of things sampled should lie. It is therefore proposed to make some more assumptions which, for specific cases, one knows are much more nearly in accordance with fact. By comparing a posteriori probabilities obtained under these more nearly correct assumptions with the probability obtained under a priori equal likelihood, one can get some idea of the limits within which the general assumption can be used as a practical working basis. These comparisons will show that if n is reasonably large (of the order of 100 or greater) and if P is large (of the order of 0.90 or greater), then the distribution curve for the fraction effective is so sharp and steep that a rather wide divergence between assumption and fact, even to the extent of assuming equal likelihood, generally makes very little difference in results. Therefore, under such conditions, this procedure is of practical value. For P less than 0.90, of course, the sample size should be larger or else one takes increasing risk both of lack of precision and of bias.

Suppose that one has taken the sample ($n = 10$, $c = 2$) from an unknown lot of articles, but from long previous experience one knows that this type of article, although severely defective, practically always has a P between 0.60 and 1.00, and that the average value of P is about 0.83. Given such experience, the actual distribution could be determined with a high degree of approximation by mathematical means,[10] but by way of illustration let it be assumed that it is approximately as shown in Fig. A2.

[10] A well-known procedure consists of employing several terms of the Gram-Charlier series. The first term involves only the parameters associated with normal law, viz., the average, $\bar{X}$, and standard deviation, σ. As few as 100 observations may yield fairly precise estimates of these parameters. The second term involves the measure of skewness or lopsidedness of the curve, β_1. Something of the order of 500 or more observations should be employed in estimating this parameter. If consideration must be given to the flatness or peakness of the curve, the third term of the series should be employed, which involves the measure of kurtosis, β_2. In this estimate something of the order of 5000 observations should be employed. It is therefore apparent that empirical determinations of distributions may, in practice, involve considerable difficulty.

The present assumption states that the lot sampled is one drawn at random from a large stock of lots distributed as shown in Fig. A2. The probability that a lot of any given fraction effective is the lot sampled is calculated as illustrated in Table A2; viz., the probability that the lot would produce the sample is multiplied by the frequency with which that kind of lot occurs, and this product is divided by the sum of like products for all possible lots. Carrying out this process on a basis of lots at 1% apart yields the solid curve

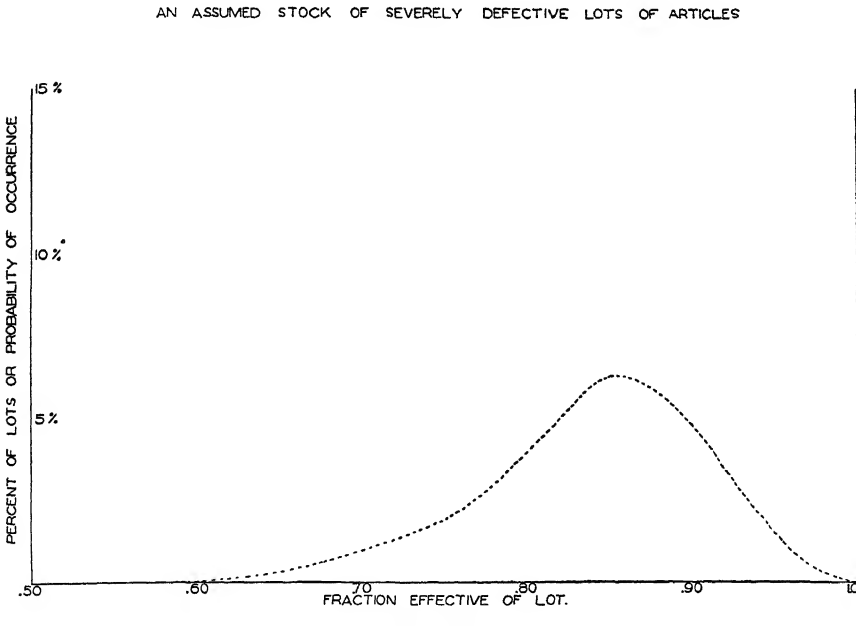


FIG. A2.

of Fig. A3. Obviously, the information is less indefinite than that yielded by Fig. A2. To improve the comparison the two figures are plotted together to the same scale in Fig. A3.

To complete the picture of the effect of the a priori distribution in connection with very small samples, let one assume that the lot came from a stock of lots which is known to be only moderately defective, so that practically all lots are between $P = 0.80$ and $P = 1.00$, with mean at 0.925. The a priori distribution or existence probability curve together with the corresponding a posteriori probability curve are shown in Fig. A4.

PROBABILITY THAT THE SAMPLE $n = 10$, $c = 2$ CAME FROM ANY LOT, ASSUMING LOTS BETWEEN 0.60-0.99

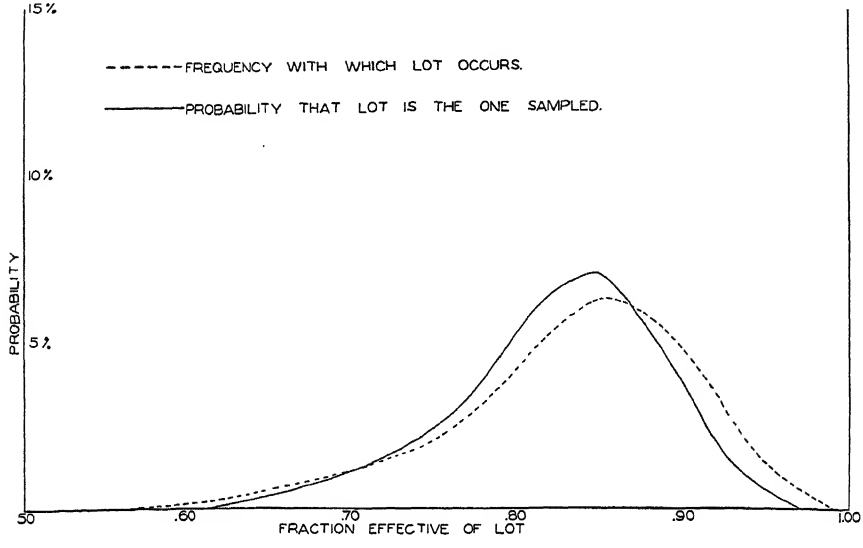


FIG. A3.

PROBABILITY THAT THE SAMPLE $n = 10$, $c = 2$ CAME FROM ANY LOT, ASSUMING LOTS BETWEEN 0.80-1.00

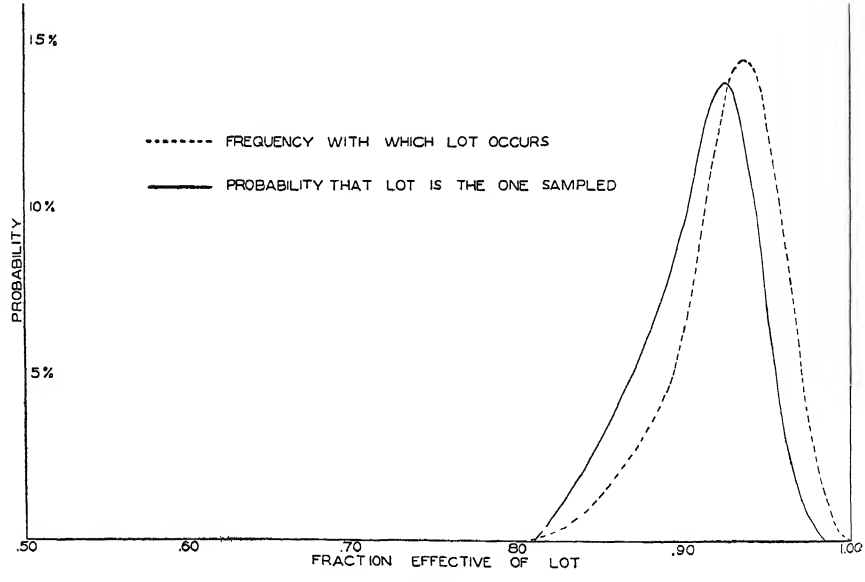


FIG. A4.

To summarize, all six curves are shown in Fig. A5. The downward-pointing arrows on the three respective probability curves indicate the middle value of the lot fraction effective which may have produced the sample. It divides the area under the curve into two equal parts. It is the median, or value such that in repeated trials the estimate would be too high just as frequently as too low.[11] The upward-pointing arrows mark limits which are offered for the purpose of estimating the precision of the median. They cut off tails

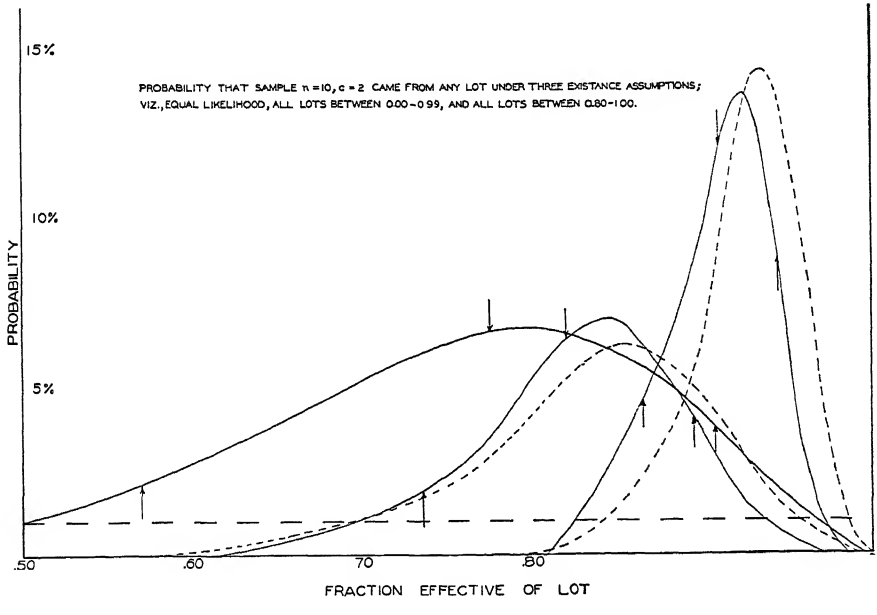


FIG. A5.

of one-tenth the area of the curve. Thus, whereas the median yields the estimate of the lot, the upward-pointing arrows show two limits such that (based on the hypothesis and the evidence presented) the probability is 0.9 that the fraction effective of the lot sampled is less than one limit and the probability is 0.9 that it is greater than the other. The following deductions can be drawn from the curves:

[11] It should also be noted that in the binomial distribution the median is very close to the mean.

Assumption of distribution of actual fractions effective	*Estimate of fraction effective of lot*	*Limits of fraction effective of lot (changes 9 out of 10 for each limit)*
Equal likelihood	0.76	0.57–0.92
Lots between 0.60 and 0.99 (mean at 0.83)	0.83	0.73–0.89
Lots between 0.80 and 1.00 (mean at 0.925)	0.91	0.81–0.92

From this tabulation one sees that, the narrower the a priori range of possible values for the fraction effective, the more definite the probable conclusions about the individual lot from which the small sample was taken. On these grounds it appears desirable to have as accurate an a priori assumption as possible, and to hope that such an assumption will have a very narrow range. However, one danger immediately suggests itself. The estimated value of the lot fraction effective is drawn toward the mean value of the assumption; and the more narrow and peaked the assumption, the more it is drawn. For example, in the assumption of equal likelihood, the estimate is 0.763, a slight shift from the observed fraction effective of 0.80 toward the equal likelihood mean of 0.50. In the assumption of lots between 0.60 to 0.99, the estimate is 0.83, a shift toward the assumption mean of 0.83. In the assumption of lots between 0.80–0.99, the estimate is 0.91, a very pronounced shift toward the assumption mean of 0.925. Hence, when the assumption is narrow and peaked, it exerts a very powerful influence on the estimate, which is well and good if it should be that way, but which may well lead to quite false conclusions if there is an error in the assumption. However, before judging too closely, the effect of larger sample size should be investigated.

Effect of sample size on a posteriori probability. With the foregoing paragraphs as a background, the effect of sample size on accuracy of the estimate can be presented swiftly. If the sample had consisted of 50 articles and if the sample had been found to be 0.80 fraction effective, the curves for the probability that the sample came from any lot under the three respective assumptions would be as shown in Fig. A6. Here the estimates of the fraction effective of the lot are brought much closer together, viz., 0.79, 0.81, and 0.87, respectively, for equal likelihood, lots between 0.60 and 0.99, and lots between 0.80 and 1.00. This is because of the increased influence of the sample, due to its larger size, which tends to drag all estimates toward itself, viz., observed sample fraction effective 0.80.

This is a wholesome influence, since, in general, the sample is a more intimate witness of the lot value than the mean associated with the lots; but it is purchased at the cost of increased samples.

Had the sample consisted of 400 items, and had the sample been 0.80 fraction effective, the curves would have been as shown in Fig. A7. Here the respective estimates are 0.80, 0.81, and 0.835. This figure illustrates the fact that, even though the sample is large, it is rather important that not the wrong assumption be chosen. If

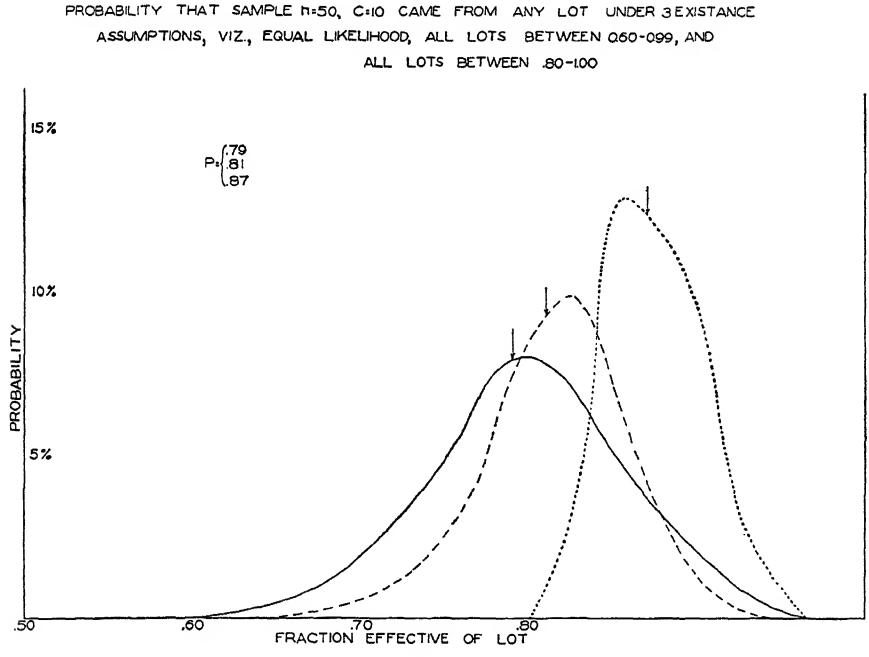


FIG. A6.

the lot sampled really has a fraction effective of about 0.80 or 0.81, as some lots under the third assumption must have, the chances are rather small of its being rated at that figure, unless the sample happens to give considerably more than the average number of defectives. Hence, the danger of the narrow peaked assumption is obvious. On the other hand, if the assumption is moderately widened, as in the case of the 0.60–0.99 assumption (see Fig. A7), the width of indefinite range is very little improved over that of equal likelihood, when the sample is large.

Effect of the fraction effective on a posteriori probability. The preceding charts have already partially illustrated the effect of P on the estimate. To illustrate further the effect of P on accuracy and its relationship to the assumption, Fig. A8 has been drawn. In this instance the sample is still 400, but the sample proved to be 0.95 fraction effective. With these values of P and n, all three assumptions give practically the same result, viz., 0.95, 0.95+ and 0.95−.

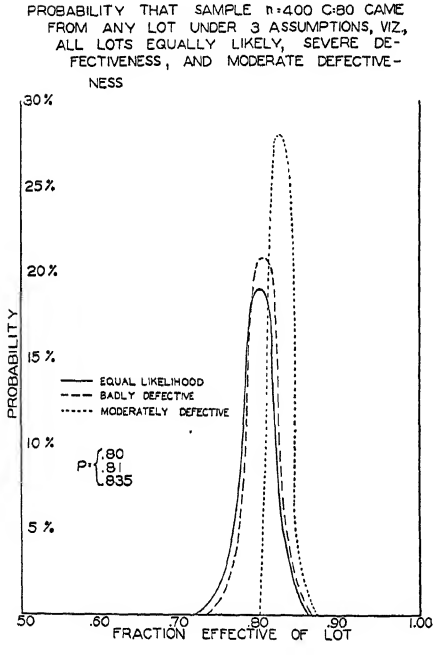


FIG. A7.

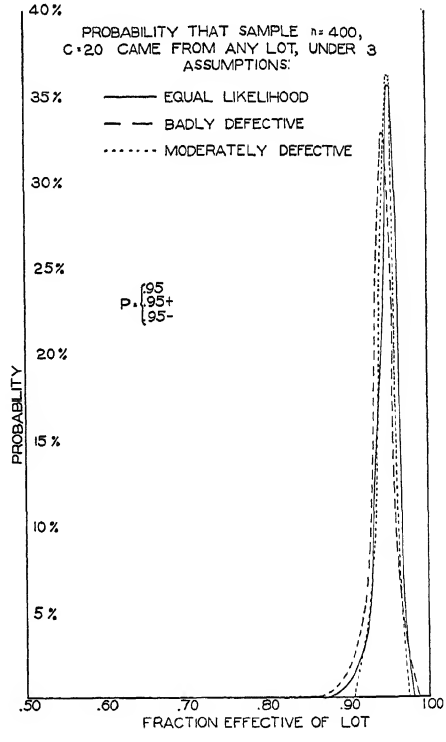


FIG. A8.

When increase in data gives little increase in knowledge. It may sometimes occur that quite large samples can be made available without great cost or effort. Let it be observed that, although the discussion has related to lots, all remarks made thus far would be equally true regarding the fraction defective (or effective) of any aggregation whatsoever irrespective of its composition and lack of homogeneity, because the discussion has not required, made, or implied any assumptions regarding homogeneity. It has also been

observed that minimum sample sizes of considerable magnitude are unavoidable and that, incidentally, these large sample sizes may justify the assumption of a priori equal likelihood. Is there also some maximum sample size taken from an aggregation beyond which it does not pay to go?

RELATIONSHIP BETWEEN SAMPLE FRACTION EFFECTIVE AND PROBABLE LOT FRACTION EFFECTIVE.

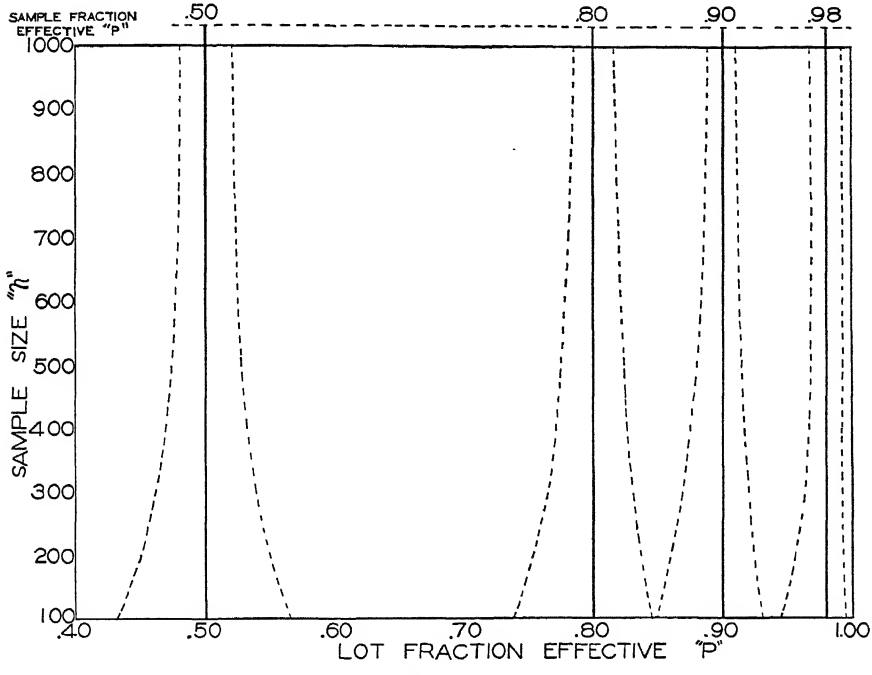


FIG. A9.

Figure A9 has been constructed to show the probable lot fraction effective for four sample fractions effective; lot, in this case meaning lot, grand lot, or aggregation of grand lots. The solid lines show the middle value, under the a priori assumption of equal likelihood, and the dotted lines show the 0.9 probability limits. Thus, if a sample of 500 were taken from a grand lot and found to be 0.98 fraction effective, its probable lot fraction effective is almost 0.98 and the chances are 0.9 that it is less than about 0.987 and the chances are 0.9 that it is greater than about 0.97. If the sample were increased to

1000, the accuracy would be only a little better, but a material decrease in sample size below 500 results in markedly less accuracy. Hence, even if a superpopulation of several grand lots were sampled with a view to determining the a priori distribution, even the mean of this superpopulation would not be known much more accurately than the grand-lot mean. It therefore appears that a practical solution of the problem is to sample grand lots, not lots or superpopulations, and to use the procedure outlined in Chapters III and IV, when this is possible. If this procedure is not possible, then careful consideration should be given to the effects of the assumption, the sample size, and P.

Summary of the effects of assumption, sample size, and fraction effective. If the lots sampled are of high fraction effective (0.90 to 1.0) and the sample is large, the a priori assumption regarding the universe of lots sampled is of little consequence, provided only that a too peaked distribution curve is not assumed. If the lots are of poorer quality, even though the sample is large, the a priori assumption of the actual distribution is significant. If the sample is small, the assumption is always of importance, but it is not practical to make reliable estimates of lot fractions effective from very small samples under any arrangement of a priori assumptions.

Estimation of lot fraction effective without any a priori assumption. Despite the demonstration of the fact that the assumption of a priori equal likelihood will have little, if any, ill effect when used in an experimentally limited range, some persons may still prefer some other means [12] of estimation. It is therefore suggested that one consider constructing a chart for estimating the fraction defective, Q, on the basis of selecting a Q such that the probability that such a lot would produce more than the observed number of defectives, c, in the random sample of n is equal to the probability that the same lot would produce less than the observed number of defectives, c, in the random sample of n. Such a value of Q (which could also be designated Q_M) would be a median value of Q, since it is obvious that, if the experiment were repeated a large number of times, the number of times that the estimate of Q was greater than true Q would approximate the number of times that the estimate was less than true Q.

By the use of the incomplete beta-function ratio, which is dis-

[12] See "On the Problem of the Most Efficient Tests of Statistical Hypotheses," J. Neyman, *Phil. Trans., Royal Society (London)*, Vol. CCXXXI-A, pp. 289-337, London, Feb. 16, 1933.

cussed in Appendix B, the construction of such a chart is readily possible.

The probability that a lot Q would produce $c + 1$ or more defectives in a sample of n is the incomplete beta-function ratio,

$$I_Q(c + 1, n - c).$$

The probability that the lot Q would produce $c - 1$ or less defectives is 1 — the probability that it would produce c or more defectives, or

$$1 - I_Q(c, n - c + 1).$$

In order to solve for a Q which will make these two probabilities equal, let

$$I_Q(c + 1, n - c) = 1 - I_Q(c, n - c + 1)$$

or

$$I_Q(c + 1, n - c) + I_Q(c, n - c + 1) = 1.$$

The solution now reduces to fixing pairs of values of c and n, and then finding from the *Tables of the Incomplete Beta-Function Ratio* a value of Q which satisfies the above equation. For example, let $c = 1$, and $n = 4$. Then $I_Q(2, 3) + I_Q(1, 4) = 1$. The tables list $I_Q(p, q)$, when $p \geq q$. Hence one must read

$$I_{1-Q}(3, 2) \text{ and } I_{1-Q}(4, 1).$$

Ordinarily, we would subtract each from 1, add, equate to unity, and solve for Q.

However, if $x + y = 1$, then

$$(1 - x) + (1 - y) = 1.$$

Hence, it is not necessary to subtract from 1 before adding.

For $1 - Q = 0.73$, i.e., for $Q = 0.27$:

$$I_{1-Q}(3, 2) + I_{1-Q}(4, 1) = 0.2840 + 0.7041 = 0.9881.$$

For $Q = 0.26$, $1 - Q = 0.74$:

$$I_{1-Q}(3, 2) + I_{1-Q}(4, 1) = 0.2999 + 0.7213 = 1.0212.$$

For $Q = 0.25$, $1 - Q = 0.75$:

$$I_{1-Q}(3, 2) + I_{1-Q}(4, 1) = 0.3164 + 0.7383 = 1.0547.$$

Interpolating, the value of Q which yields the sum equal to 1 is 0.266. From the Q_M Chart one reads 0.314 for $n = 4$, $c = 1$.

A summary of results for this and other values obtained in similar manner is shown in Table A3.

TABLE A3

COMPARISON OF ESTIMATES OF Q_M WITH AND WITHOUT AN A PRIORI ASSUMPTION

(c, n)	Q to give c's greater than observed c and less than observed c with like frequency	Q by a priori assumption of equal likelihood
$c = 1, n = 4$	0.2660	0.3140
$c = 1, n = 11$	0.1010	0.1360
$c = 2, n = 25$	0.0851	0.1010
$c = 12, n = 25$	0.4800	0.4800
$c = 1, n = 50$	0.225	0.0325
$c = 10, n = 50$	0.2020	0.2080
$c = 20, n = 50$	0.4000	0.4040
$c = 1, n = 5$	0.2161	0.2600
$c = 2, n = 5$	0.4056	0.4160

It will be seen that, as n increases and as Q approaches 0.5, the difference becomes small.

If $c = 0$, then $I_Q(c, n - c + 1) = 1$ for any Q, since the probability of 0 or more failures must always be unity, regardless of Q. Hence if:

$$I_Q(c, n - c + 1) + I_Q(c + 1, n - c) = 1,$$

then

$$I_Q(c + 1, n - c) = 0.$$

The only Q which has a zero probability of yielding $c + 1 = 1$ or more failures is $Q = 0$. This can also be explained as follows:

We are looking for a Q which will produce failures greater than the observed c and less than the observed c with equal frequency; i.e., c is the median value. If $c = 0$, the only Q that will yield a median c of zero is $Q = 0$. (This is also true for an average $c = 0$ or modal $c = 0$).

The practical answer, however, is that, where $c = 0$, there are not enough samples to give a definite estimate. For example, if $c = 1$, $n = 11$, we would estimate that the Q_M is 0.101 (see above table). However, if $c = 0$, $n = 11$, we could only say that our estimate is that Q_M is less than 0.101.

In accordance with this procedure, the author entered dotted lines on a Chart $0.5 = I_Q$ for $c = 1, 5, 10, 15$, and 20. As can be seen

from Table A3, the modified chart showed that, as the sample size increases, and for moderate defectiveness, the difference in the estimated Q_M under the two procedures is small—so small, in fact, as to be negligible under certain conditions for n greater than about 25. It might appear preferable to use curves predicated on no a priori assumption. However, since the latter curves do not differ appreciably from the former for any except relatively small samples (which are at best unreliable), since the latter are difficult to compute, and since in sampling work the former give estimates on the side of safety, the curves as shown on Chart $0.5 = I_Q$ are recommended.

It is obvious that like reasoning could be used for obtaining estimates of $Q_{L0.1}$ and $Q_{U0.9}$ without any a priori assumption; i.e., one could look for a lot of fraction defective $Q_{L0.1}$ such that the probability would be 0.9 that less than c defectives would occur in a random sample of n. In this case one has $I_Q(c, n - c + 1) = 0.1$. This is a regular incomplete beta-function ratio and involves no cut-and-try solution. It is notable that it is the same solution which one would obtain by a posteriori probability from Chart $0.1 = I_Q$, except that one fails to increase n by one. However, in the light of the discussion of these upper and lower limits as given under "The Logic of Probable Judgments," in Chapter II, there appears to be no reason for such procedure.

APPENDIX B

ON RELATIONSHIPS OF THE INCOMPLETE BETA-FUNCTION RATIO AND THE CALCULATION OF THE CHARTS FOR ITS SOLUTION

General discussion. The complete beta function,

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx,$$

and the incomplete beta-function,

$$B_x(p,q) = \int_0^x x^{p-1}(1-x)^{q-1}dx$$

are well known.[1] The incomplete beta-function ratio,

$$I_x(p,q) = \frac{B_x(p,q)}{B(p,q)},$$

is less common. This function is available in tabulated form[2] with values of the probability given to seven decimal places and values of p and q up to 50. However, in the application of statistical methods to engineering work one desires the function for values as high as $q = 500$, although one is, in general, not interested in precision in the probability beyond two or three decimal places. Since the calculation of the published tables required the work of a staff of approximately eleven workers for eight years, the necessity of a less laborious method of approximate evaluation is obvious, in order to extend the function to the desired range even for a few chosen probabilities.

It should also be remarked that the published tables are for various statistical uses and yield the probability as a function of p and q, whereas in engineering work one generally wishes values of p and q for a chosen value of the probability. Such a type of solution from the tables is quite too time-consuming for the engineer who wishes to use statistical methods as an auxiliary tool in accomplishing his

[1] See any textbook on advanced calculus.

[2] *Tables of the Incomplete Beta-Function*, edited by Karl Pearson, F. R. S., The Biometrika Office, University College, London, 1934.

professional ends. By presenting the function in graphical form, the type of solution which the engineer desires is instantly available without any impairment of other uses of the function.

In this book, and in the charts supplied herewith, p has been replaced by c, and q by $n - c + 1$, because of the clarity of this notation in indicating the summation of terms beyond the cth in the binomial $[(1 - Q) + Q)]^n$, which is the principal practical use that is herein made of the function.

The use of the function in connection with a posteriori probability is less important, but it is believed that the exposition of its engineering uses and the procedure by which it was evaluated will proceed more smoothly if the proof of its relationship to a posteriori probability (under the assumption of a priori equal likelihood) is given first.

A posteriori probability. Having observed a certain number c defective articles in a relatively small sample of n articles drawn at random from a large lot, what estimate should one make of the proportion Q of defective articles in the lot? Without attempting a complete analysis, which would require the repetition of much detailed matter available in textbooks on probability, it may be stated that the problem presented is a characteristic one in a posteriori probability. A brief discussion of a posteriori probability is given in Appendix A. The formula for the solution of the problem is well known.

The probability $W(Q_L,Q_U)$ that the true proportion of failures in the lot lies between the limits Q_L and Q_U is by the Laplacian generalization of Bayes' formula,

$$W(Q_L,Q_U) = \frac{\int_{Q_L}^{Q_U} W(x) \cdot x^c(1 - x)^{n-c} dx}{\int_0^1 W(x) \cdot x^c(1 - x)^{n-c} dx}, \qquad (1)$$

where $W(x)$ is the a priori probability that a lot with the proportion of defectives $Q = x$ existed; and $x^c(1 - x)^{n-c}$ is the simple productive probability by the binomial expansion [3] that such a lot would produce the observed sample.[4]

[3] The coefficient $\frac{n!}{c!(n - c)!}$ is omitted in both numerator and denominator, as they cancel out.

[4] One should compare the logical parallelism of meaning of equation 1 with the simple exposition of a posteriori probability given in Appendix A.

This formula, however, is incapable of solution without some definite knowledge or assumption regarding the analytical form of $W(x)$. For lack of knowledge regarding $W(x)$, and because such assumption yields results of wide practical application, it is proposed to assume that $W(x)$ is a constant; i.e., that $W(x)$ is independent of x. The practicality of this assumption of equal likelihood is discussed in Appendix A. With the aid of this assumption equation 1 can be reduced to [5]

$$W(Q_L,Q_U) = \frac{\int_{Q_L}^{Q_U} x^c(1-x)^{n-c}dx}{\int_0^1 x^c(1-x)^{n-c}dx} = \frac{(n+1)!}{c!\,(n-c)!}\int_{Q_L}^{Q_U} x^c(1-x)^{n-c}dx. \tag{2}$$

For the probability that Q is not greater than some arbitrarily assigned value Q_U, put Q_L equal to zero, and

$$W(0,Q_U) = \frac{(n+1)!}{c!(n-c)!}\int_0^{Q_U} x^c(1-x)^{n-c}dx. \tag{3}$$

There is no convenient method of solution of equation 3 over a wide range of values of n and c. There are two well-known methods of approximate solution. Laplace transformed the equation so as to permit an approximate solution by means of the normal curve, subject to the limitation that $\frac{(Q-P)^2}{nPQ}$ be small. Another transformation into the Poisson binomial expansion limit is also possible.[6] A part of this treatment will be used subsequently. This method presupposes that Q is very small and n very large. Thus neither method is suitable for small values of n. It is interesting, however, to note that both these theorems may be regarded as special cases of the

[5] Integral tables give $\int_0^1 x^l(1-x)^m dx = \frac{l!m!}{(l+m+1)!}$, or $\int_0^1 x^c(1-x)^{n-c}\,dx$ may be recognized as the complete beta function and treated as follows

$$B(c+1,n-c+1) = \frac{\Gamma(c+1)\Gamma(n-c+1)}{\Gamma[(c+1)+(n-c+1)]}$$

$$= \frac{\Gamma(c+1)\Gamma(n-c+1)}{\Gamma(n+2)} = \frac{c!(n-c)!}{(n+1)!}.$$

Also, see latter part of note 8, page 183.

[6] "Application to the Binomial Summation of a Laplacian Method for the Evaluation of Definite Integrals," E. C. Molina, *Bell System Technical Journal*, January, 1929.

incomplete beta-function ratio expressed in equation 3. In order, therefore, to make practical use of this expression, it becomes desirable so to transform it that reasonably accurate solutions can be readily obtained for all ranges of c, n, and x.

The general point binomial. An article by George A. Campbell [7] discusses the general point binomial [8]

$$P(c, n, a) = \frac{n!}{(c-1)!\,(n-c)!}\int_0^{a/n} x^{c-1}(1-x)^{n-c}dx, \qquad (4)$$

where P is the a priori probability of at least c defectives in n trials, when the average number is known to be a. The article includes tables and charts for the complete solution over a wide range of the special case $P(c, \infty, a)$ which is the Poisson exponential binomial limit. The special case of $n = \infty$ would not be very helpful in the present problem, but Campbell offered in a footnote an approximation formula by which values of $P(c, \infty, a)$ can be altered to apply to $P(c, n, a)$, where n is any value. The nature of the approximation will be discussed subsequently.

[7] "Probability Curves Showing Poisson's Exponential Summation," *Bell System Technical Journal*, January, 1923.

[8] Ordinarily, one would think of the a priori probability of at least c defectives out of n trials when the average number is a as being $P(c, n, a)$ = sum of the last $(n - c + 1)$ terms of the binomial $\left[\left(1 - \frac{a}{n}\right) + \frac{a}{n}\right]^n$ or

$$\sum_{x=c}^{x=n} \frac{n}{(n-x)!\,x!}\left(1-\frac{a}{n}\right)^{n-x}\left(\frac{a}{n}\right)^x = \frac{n!}{(n-c)!\,c!}\left(1-\frac{a}{n}\right)^{n-c}\left(\frac{a}{n}\right)^c$$
$$+ \frac{n!}{(n-c-1)!\,(c+1)!}\left(1-\frac{a}{n}\right)^{n-c-1}\left(\frac{a}{n}\right)^{c+1}, + \cdots, \left(\frac{a}{n}\right)^n.$$

Formula 4 can be transformed into this form. If formula 4 is integrated by parts, using the following formulas from standard integral tables,

$$\int x^{m-1}z^p dx = \frac{x^m z^p}{m+p} \frac{ap}{m+p}\int x^{m-1}z^{p-1}dx,$$

$$\int x^{m-1}z^p dx = \frac{x^m z^{p+1}}{am} - \frac{b(m+p+1)}{am}\int x^m z$$

where $z = a + bx$, and using the first formula and then the second alternately, it will be found that the result of the integration is identical to the last $c + 1$ terms of the binomial expansion indicated above.

Equation 4 may also be recognized as the incomplete beta-function ratio

$$I_x(c, n-c+1) = \frac{\mathrm{B}_x(c, n-c+1)}{\mathrm{B}(c, n-c+1)}.$$

Transformation from a posteriori to a priori probability. In equation 3, substitute $(n' - 1)$ for n and $(c' - 1)$ for c, and observe that a change in the constants results in no change in the limits of integration. Then

$$W(0,Q_U) = \frac{n'!}{(c' - 1)!\,(n' - c')}\int_0^{Q_U} x^{c'-1}(1 - x)^{n'-c'}dx.$$

This equation is exactly similar in form to Campbell's equation 4 except for the limits of integration. However, if one ignores the fact that the equation represents an a posteriori probability, and regards it as an a priori probability like equation 4, then Q, the proportion of defectives will of necessity be equal to the average number of defectives divided by the sample size, in this case, a'/n'. Hence,

$$W(0,Q_U) = P(c', n', a') = \frac{n'!}{(c'-1)!(n'-c')!}\int_0^{a'/n'} x^{c-1}(1-x)^{n'-c'}dx. \quad (5)$$

The form of equation 5 is now the same as Campbell's equation 4; hence, by observing the relationship in the meaning of the terms, the problem in a posteriori probability can be transformed into one in a priori probability [9] and the curves and tables of Campbell's article can be used with extremely close approximation in evaluating the formula under consideration; that is to say, if c, n, and Q_U are the terms of equation 3, then Campbell's charts must be entered using $c + 1$, $n + 1$, and $Q_U/(n + 1)$. Specifically, the meaning of this relationship can be expressed as follows:

The a posteriori probability that a large lot contains a proportion of defectives not greater than an arbitrarily assigned value Q_U when c defectives have resulted from a sample of n articles is equal to the a priori probability of obtaining at least $c + 1$ defectives in a random sample of $n + 1$ articles taken from a lot in which the proportion of defectives is Q_U.

This relationship is as expected since Paul P. Coggins proved that the a posteriori probability that a universe of N items contains not more than X defectives when c defectives have resulted from a

[9] Mr. Paul P. Coggins in "Some General Results of Elementary Sampling Theory for Engineering Use," *Bell System Technical Journal*, January, 1928, accomplishes a like transformation for the probability $W(X,X_2)$, having observed c defectives in a sample of n, taken from a lot of size N. He develops charts for a wide range of lots, samples, and several probabilities, and takes cognizance of lot size. Charts of a more generalized character and of maximum convenience in use are sought in the present case.

random sample of n items is equal to the a priori probability of obtaining at least $c + 1$ defectives in a random sample of $n + 1$ items from a universe of $N + 1$ items of which exactly $X + 1$ are defective.

The foregoing discussion explains the meaning of the notes on the charts I_Q which provide for the reading of both a priori and a posteriori probabilities from the same chart. It also explains the mathematical basis of the charts. To complete the discussion, the approximation method for n finite should be explained.

Campbell's tables and charts give values for P, c, and a. For example, for $P = 0.5$, $c = 1$, the corresponding value of a is 0.6931. This means that the probability is 0.5 of obtaining at least 1 failure out of a very large sample when the averaging number of failures in samples of that size is 0.6931. The sample size is pointedly omitted, for theoretically it should be infinity. If it is assumed that n is 1000, then the above data means that the probability is 0.5 of obtaining at least 1 defective in a sample of 1000, when Q (the average number divided by n) is approximately 0.0006931. The answer is approximate, because n is no longer $\propto$. If one assumes n to be 50, of course, the answer

$$Q = \frac{a}{n} = \frac{0.6931}{50} = 0.013862$$

is less proximate as n now departs further from infinity.

However, Campbell gave the following relationship:[10]

$$\frac{a(c,n,P) - a(c,\propto,P)}{a(c,\propto,P)} = An^{-1} + (\tfrac{1}{12})[14A^2 + (3a+2)A + a]n^{-2} + \cdots,$$

where $A = (\frac{1}{2})(c - a - 1)$ and $a = (a,\propto,P)$.

Now, for the finite sample size $n = 50$, it is easy to find the approximate a.

$$\frac{a_{50} - a}{a} = An^{-1} + (\tfrac{1}{12})[14A^2 + (3a + 2)A + a]n^{-2},$$

$$a = 0.6931,\ A = (\tfrac{1}{2})(1 - 0.6931 - 1) = -0.34655.$$

[10] I am indebted to Mr. R. M. Foster of the Bell Telephone Laboratories for delving into Mr. Campbell's original material, which he left at the Bell Telephone Laboratories at the time of his retirement, and explaining to me, at Mr. Campbell's request, the derivation of this series. Additional terms for closer approximation are not theoretically difficult to derive, but are laborious, and the next term is too complicated to be of service as an easy method of approximation.

Therefore

$$a_{50} = 0.6931 - \frac{(0.6931 \times 0.34655)}{50}$$

$$+\left(\frac{1}{12}\right)\left[\frac{14(0.34655)^2 - (3 \times 0.6931 + 2)(0.34655) + 0.6931}{50 \times 50}\right](0.6931),$$

$$= 0.6931 - 0.00480 + 0.00002,$$

$$= 0.6883,$$

$$Q = \frac{a}{n} = \frac{0.6883}{50} = 0.01377.$$

Now, carefully noting the meaning of the terms, one recalls that n and c in equation 4 are each one greater than n and c in equation 3. Therefore, the point on Chart 0.5 = I_Q for which the solution has just been made is $c = 0$, $n = 49$, $Q = 0.01377$. This is the only chart on an a posteriori basis. In working with the charts for 0.005, 0.1, 0.9, and 0.995 probabilities, it is not necessary to subtract 1 from c and n. In this manner solutions were made for 9000 points from which the charts were plotted. Additional corrections were made by the method of simple combinations, where the formula did not yield the desired accuracy.[11] In a few cases when n was greater than 400 and Q was exactly 0.50, check points could be obtained from the normal distribution.

The accuracy and precision associated with the charts. Except for very high and very low ranges of Q, the data for the charts were calculated to three significant figures. The plots were made to a large scale, the curves carefully smoothed, and the charts finally photostated down to reduced size. The grid of the charts is especially designed to spread the curves over the surface and to keep the precision in significant figures as nearly uniform as possible. Actually, the calculations of the points were carried to four figures and then rounded off; and this precision, together with that lent by graphical smoothing on a large scale, tends to give the charts con-

[11] Since $$\frac{n!}{(c-1)!\,(n-c)!}\int_0^{a/n} x^{c-1}(1-x)^{n-c}\,dx$$ equals the sum of the last $n - c + 1$ terms of the binomial $\left[\left(1 - \frac{a}{n}\right) + \frac{a}{n}\right]^n$, evaluation of equation 5 is possible but highly laborious by direct computation even using tables of the logarithm of the factorial. Other methods are given in *Tables for Statisticians and Biometricians*, Cambridge University Press, London.

siderably greater precision than that to which they can be read. It should be noted, therefore, that for practical purposes the accuracy [12] of results obtained from the charts is, in the case of a priori probabilities, a function of the accuracy of the data with which they are entered. In the case of a posteriori probabilities, it is a function both of the accuracy of the data and of the appropriateness of the assumption of a priori equal likelihood.

[12] The distinction between accuracy and precision is important to the engineer. Accuracy relates to deviation from the true value; precision is associated with mere reproducibility. For example, a one foot scale graduated to fiftieths of an inch might be regarded as a very precise measuring instrument. However, if the same scale (perhaps due to some accident), contained an initial error of $\frac{1}{16}$ inch, its accuracy would be very poor. An enlightening discussion is contained in *Statistical Method from the Viewpoint of Quality Control*, W. A. Shewhart, The Graduate School, Department of Agriculture, Washington, D. C., 1939.

APPENDIX C

SAMPLE ILLUSTRATION OF A SIMPLE, WORKING SYSTEM OF QUALITY CONTROL

With Various Control Chart Factors for the Use of Range

Introduction. When it is decided that it is desirable to institute a system of statistical quality control in an organization, two very practical problems arise: (*a*) the design of a system which is adapted to the technical and mathematical abilities of existing plant personnel, and (*b*) the procedure for gathering initial data for setting control limits.

These problems were met in somewhat acute form when it was decided to institute a system of statistical methods for quality control at Picatinny Arsenal, Dover, N. J. No men were available who had more than a high-school education, and no records suitable for computing control limits were extant. Accordingly a system was devised which is so simple that it was readily administered by existing personnel with no preliminary instruction. As an acid test of the simplicity, small pamphlets precisely as reproduced in this Appendix were issued to the foremen and inspectors a few days prior to the order directing that the system be put into effect. The system worked from the start with only a few questions from operating personnel.

Prior to instituting the system, a careful study was made of the processes it was to control. Trials were made of types of systems, using various sizes of subgroups of samples. In this process, parallel control charts were kept based both on range and standard deviation, with no relaxation in the existing inspection system. These measures were taken to insure a properly working system at the beginning of its establishment.

Economic advantages of the system. Prior to instituting the system, the management of the Arsenal sharply raised the question of how many additional inspectors would have to be hired to adminis-

trate this precise and scientific system of quality control. The preliminary surveys for the institution of the system had been made, and the management was informed that no new personnel need be employed: on the contrary, over half the existing force of inspectors could be made available to the management to overcome an existing shortage of foremen in other departments of the Arsenal not concerned with the new system.

Despite the marked reduction in the cost of inspection, the inspection function was performed far better than it had ever been before, the quality of product was much improved both in uniformity and aimed-at level, and costs were reduced. The outstanding example of improvement of quality consisted in holding a quality characteristic (the weight of explosive in a certain piece part) to $\bar{X} \pm \frac{1}{3}$ gr. (a standard deviation of 0.013), on which the specification was $\bar{X} \pm 2$ gr., and about which repeated complaints had been previously received from the assembling arsenal for violations of tolerance limits. The outstanding reduction of cost of an operation consisted of a reduction of 85% which was due (*a*) in part to reduced cost of inspection, and (*b*) in part to reduced cost of production, because the quality control technique made possible the employment of statistically tested loading machines [1] which would hold the product well within tolerance limits. In no instance did costs increase.

It is not suggested that the system as shown in the illustration is of general application.[2] A system should be carefully fitted to the engineering principles governing the products in question. This system is offered merely as a guide, and especially as an illustration of the simplicity to which the administration of statistical methods can be reduced. In the illustration the older notation w is used for range instead of R. An extensive tabulation of various limits for use in systems of various choice together with a brief comment on their derivation is offered at the end of this Appendix.

As occasion may arise when the reader may find a general treatment more useful than a specific example, an outline of the considerations involved in the design of a simple quality control procedure of this kind is furnished.

[1] For a method of testing machines statistically, see L. E. Simon, "Deviations in Product Prove Machine Performance," *Product Engineering*, December, 1936.

[2] It should be noted that par. 18 of the sample procedure does not provide a control chart in the usual sense, but was merely a special procedure for defecting a future product which appeared to be different from the previous kind.

Check List of Steps in a Simple Quality Control Procedure [3]

1. Attain both statistical and engineering competence in design of quality control procedure.
   - *a.* Careful study of the production lines by the statistician.
   - *b.* Collaboration of the statistician with an industrial engineer who has some statistical knowledge.
2. Choose statistical techniques which are simple and economic.
   - *a.* Adjust sample sizes so as to employ a statistic which is simple of calculation, with perhaps more mechanical effort, rather than a technically more efficient one.
   - *b.* Weigh the economic aspects of increased sampling versus more efficient statistics.
   - *c.* Consider the physical and engineering interpretation of the techniques, and their appeal to the engineer, industrialist, or executive.
3. Keep full emphasis on the economic viewpoint.
   - *a.* Consider statistical aids for the better attainment of the economic aims of the process.
     - (1) Larger volume.
     - (2) Greater uniformity of product.
     - (3) Reduction of cost of inspection.
     - (4) Reduction in wastage.
     - (5) Avoidance of trouble.
     - (6) Authentic record of the quality of product.
   - *b.* Avoid the temptation of attaining basic scientific knowledge which is not of immediate economic value; this can be obtained after the statistical method is working.
   - *c.* Try to make provision for the employment of personnel displaced by the more efficient process.
     - (1) Tactfully suggest their employment in other parts of the plant.
     - (2) Tactfully suggest increased volume of production.
     - (3) Be cautious about intruding on the logical prerogatives of management.
4. Publish the Quality Control Procedure in clear written form, signed by the management.
   - *a.* Alter the existing process and personnel as little as possible without sacrifice of fundamental principles.
     - (1) Fit the statistical techniques to the existing process and engineering considerations.
     - (2) Devise methods that can be administered by existing personnel in as great a measure as practicable.
   - *b.* Reduce the procedure to a simple set of functions, steps, and consequent actions each of which is the duty of the incumbent of a designated position.

[3] Extracted from an address entitled, "On the Initiation of Statistical Methods for Quality Control in Industry" delivered by the author at the 102nd Annual Meeting of the *American Statistical Association*, Chicago, 1940.

c. Employ simple language and engineering terms; avoid technical statistical terms.

d. Delegate clearly defined and routine duties of the procedure to positions or offices such as inspector, foreman, superintendent, etc., rather than to persons.

(1) Collection and posting of data may be put under foreman.
(2) The unit which makes the product should not participate in its sampling.
(3) Interpretation of record, a responsibility of the inspector.
(4) Location of trouble, a responsibility of supervising engineer.

e. Choose inspection intervals which are appropriate to the process, such as every half hour, every 100 items, every buggy load, etc.

f. Prescribe the number of articles to be inspected at each inspection interval—by whom and how.

g. Describe the exact way in which records shall be kept, who shall keep them, and provide convenient forms for that purpose.

h. Provide the administrators with multiplying factors, charts, etc., so that they will neither have to do mathematical or statistical thinking nor have to refer to any literature other than the published Quality Control Procedure, in posting the records.

i. Provide for reduction of inspection with improvement in process control by allowing competent authority to change the inspection interval.

j. Prescribe the action to be taken when points are out of control limits on the record, such as stop the process, hunt for the cause of the trouble, inform the superintendent, etc.

k. Insure ease of location of uncontrolled product by not permitting sampled increment to continue in production, until the sampling result is seen to be within limits.

l. Arm the Quality Control Procedure with a thinking clause by providing that any case not covered by routine instructions will be referred to a designated office whose incumbent is capable of competent engineering or statistical analysis.

INSTRUCTIONS for CONTROL of QUALITY of PRODUCT thru PERCENTAGE INSPECTION

GENERAL INSTRUCTIONS

1. The following procedure is designed to govern the inspection of all manufactured items on which the nature of the work performed can be measured quantitatively; e.g., weight of explosive charge in various components, explosive power of detonators in terms of weight of sand crushed, specific gravity of cast or pressed materials, burning time of fuzes, etc., except where 100% inspection is performed.

Sampling Schemes

2. The sampling scheme described herein is based on a sample of 5 items per hour, and is practical on the majority of production orders. In some instances, however, the cost of sampling may prohibit this procedure; whereas, in others, more extensive sampling may be advisable, especially at the beginning of an order. Hence, the shop inspector will submit his recommended sampling scheme to the Department Chief for approval prior to production. The procedure for other sampling schemes is covered in notes on sampling schemes.

DUTIES of the FOREMAN

3. Take a sample of five items from the assembly line each hour, day, or other period of time, as instructed by the shop inspector.

Recording Observations

4. Accurately measure each sample with regard to size, weight, explosive power, or other characteristic described by the respective drawing and specification and record the measurements in the order taken.

Computing Data

5. a. Take the sum of the five recorded measurements of the group and divide it by 5. This figure is known as the "average" or "mean" and is designated by the symbol $\bar{X}$ (bar X).

b. For each group of 5, subtract the smallest recorded measurement from the largest recorded measurement. This figure is a measure of dispersion and is commonly known as "range" or "maximum dispersion" and is designated by the symbol W_t (W sub t).

c. Table 1 shows a sample of Foreman's Data.

Plotting Data

6. a. Plot the chart described below on cross section paper. Head the chart "Control Chart for _____", (inserting the name of the item sampled), "Samples of 5" followed by the production order number. On the face of the chart indicate the lot number or batch from which the samples were taken, the approximate daily production, and the designated measurement that the item should meet, e.g., weight of charge 30.0 gr. $\pm$ 2.0 gr., per drawing 70-1-11, revised 6-20-36. See Figures 1 to 5 inclusive.

b. On the piece of cross section paper mark a horizontal scale across the top for the working days of the month; e.g., September 1st, September 2nd., etc. Ordinarily, one linear inch for each day is convenient. If the paper has eight divisions to the inch, one division will represent a working hour of the working day.

c. Mark two vertical scales on the left hand margin of the paper; one near the top for the purpose of recording the averages ($\bar{X}$'s) and one a moderate space below it for recording the ranges (W_t's).

d. Plot the observed average ($\bar{X}$) for each group of 5 (see paragraph 5 a above) opposite the vertical scale for averages (see paragraph 6 c above) and under the horizontal scale for date and hour (see paragraph 6 b above).

e. In like manner plot the observed range (W_t) (see paragraph 5 b above) opposite the vertical scale for range and under the appropriate date and hour. Data should be plotted as promptly as practicable, and at least prior to the observation of the next group of data.

Foreman's Interpretation of the Chart

7. Limits will be placed on the chart by the Shop Inspector, within which practically all points should fall (see paragraphs 10 and 13 below). If any points fall outside of these limits, call the Shop Inspector without delay.

Disposition of Charts

8. The Foreman will conspicuously post the chart in the nearest office to the place of work, while the work is in progress; and, upon the completion of the production order will forward the chart to the Department Office for file, as a record of the quality of the product.

Delegation of Duties

9. In lieu of personally performing the functions outlined in paragraphs 1 to 8 inclusive, the Foreman may designate one or more trusted assistants to do them under his supervision. Such assistant may not in any case be the workman who performs the work being sampled.

DUTIES of the SHOP INSPECTOR

Computing and Plotting Control Limits

10. a. After the data from between eight and eighty groups of five has been plotted (see paragraph 11 below), compute the average of the observed averages. This figure is designated as $\bar{\bar{X}}$ (bar bar X). Draw a heavy horizontal line on the chart for averages at the computed figure and under the hours for which the samples were taken. See Figures 1 to 5 inclusive.

b. Compute the average of the eight to eighty observed ranges. This figure is called $\bar{W}_t$ (bar W sub t). Draw a heavy horizontal line on the chart for ranges at this value and under the hours for which the samples were taken.

c. Multiply $\bar{W}_t$ by .594 and plot two heavy dotted lines on the chart for averages parallel to the heavy line at $\bar{\bar{X}}$ and located at $\bar{\bar{X}} \pm .594\ W_t$. Mark each of the lines A 0.001.

d. In like manner, plot two heavy dotted lines on the chart for ranges, one at $2.08\ \bar{W}_t$ and one at $.254\ \bar{W}_t$. Mark each of these lines D 0.005.

Judging and Interpreting of Charts

11. a. Practically no plotted values of $\bar{X}$ should fall outside the dotted limits A 0.001 (theoretically only one above and one below in a thousand). Hence, the presence of a point outside the dotted limits is a very strong indication that the general level of quality (weight of material in a component, size, strength, or other quality) is changing from time to time. The Shop Inspector will advise the foreman to investigate at once to determine if someone is doing something wrong, if some machine is functioning wrong, if a change has been made in the raw material, etc., and the Shop Inspector will also report the situation to the Department Chief without delay.

b. A significant deviation of $\bar{\bar{X}}$ from the mean value designated by the drawing or specification obviously calls for measures to bring the average of the product in closer alignment with the designated average and the Shop Inspector will advise the Foreman accordingly. The $\bar{\bar{X}}$ from 80 groups of 5 is generally so near the true value of the product sampled that for purposes of control it may be treated as such.

c. Practically no plotted values of W_t should fall outside the dotted limits D 0.005 (theoretically only 5 above and 5 below in a thousand). The presence of a point outside these limits is a strong indication that the variation in the product (lack of uniformity) is greater than it should be. The same action will be taken as outlined in a above.

d. With respect to both charts, the plotted dots should be scattered rather evenly on both sides of the central line: the greater portion should be near the central line, and only relatively few should fall near the dotted limits. Trouble can frequently be forestalled by a study of the charts. If there is a general drift of the plotted points on either chart toward the bottom limit or the top limit, a timely investigation may eliminate the cause of the drift and prevent the occurrence of a point outside the limits. In like manner the too frequent occurrence of points at a value other than in the immediate vicinity of the central value indicates erroneous observations probably due to a faulty measuring instrument, use of an instrument not sufficiently sensitive for the work involved, or bias on the part of the observer. Action same as outlined in a above.

Number of Groups on which Limits should be based

12. In the interest of accuracy, convenience, and economy of labor, it is desirable to have limits plotted on the data from eighty groups of 5 (a normal 10 working day period). However, at the beginning of a job, limits should be calculated on the first 8 plotted points, then after a total of 16 have been accumulated, then after a total of 40, and finally after 80, all preceding points included in each successive calculation. The next set of limits will be based on the next 80 points; viz. points No. 81 to No. 160 inclusive, etc.

Predicting Limits

13. The importance of these charts lies not so much in disclosing that trouble occurred yesterday, or last week as in disclosing it instantly, or before it occurs. Hence, it is most important that limits exist for the plotted points, (see paragraph 7 above), before the points are plotted. To accomplish this purpose, the Shop Inspector will, at the time he computes and plots a set of limits for a period of eight to eighty plotted points, extend these limits in light lines for the next data period. These extended limits are binding upon production for the next period during which another set of plotted points are being accumulated (see paragraph 7 above). The limits from the accumulated data will then serve as a check on these extended limits and as a basis for new extended limits. This procedure is clearly illustrated in Figures 1 to 5 inclusive. Thus, when limits are calculated as detailed in paragraph 10 and extended as detailed in this paragraph, there are always limits predicted ahead, except for the first 8 points. Even this deficiency can be supplied by taking advantage of data from a previous order and this procedure should be followed if such data are available.

Meeting Drawings and Specifications

14. The meeting of drawings and specifications (as most of them are now written) is often more a matter of engineering judgement and interpretation than of mathematical statistics. In general, the drawing or specification will state that the product (presumably meaning every item thereof) will be A $\pm$ d. Actually, there is no way of knowing if every

item falls within the limits A $\pm$ d unless every item is sampled; and, if the sampling be destructive, there is no product left. However, if the product has showed "control" during manufacture; i.e., practically no points have fallen outside the control limits; no exhibition of a pronounced drift or trend; and, if the number of plotted points be large (e.g., forty or more), then it can be said with reasonable certainty that approximately 90% of the individual items will lie between $\overline{\overline{X}} \pm .707\ \overline{W}_t$; 95% between $\overline{\overline{X}} \pm .843\ \overline{W}_t$; and 99½% between $\overline{\overline{X}} \pm 1.21\ \overline{W}_t$. (For $\overline{\overline{X}}$ and $\overline{W}_t$ see paragraphs 10 a and 10 b respectively). Upon completing each period of 80 points the Shop Inspector will note on the chart "approximately 99½% within $\overline{\overline{X}} \pm 1.21\ \overline{W}_t$", substituting for $\overline{\overline{X}}$ its numerical value and for 1.21 $\overline{W}_t$ its numerical value.

NOTES on SAMPLING SCHEMES

Time not a Factor

15. It is not necessary that the groups of 5 be taken each hour. All the rules outlined above apply with equal force if the groups of 5 be taken every half hour, every 5 minutes, day, week, or other period of time, just so long as the observations are grouped in fives. Hence, in devising sampling schemes, sampling may be increased or decreased at will by merely varying the time interval.

16. Groups of 4 can be used just as readily as groups of 5 by changing all 5's to 4's and changing constants as follows:

Paragraph 5 a - Divide by 4 instead of 5.
Paragraph 10 c - Change .594 to .750.
Paragraph 10 d - Change 2.08 to 2.26; and .254 to .185.
Paragraph 14 - Change .707 to .798; .843 to .952; and 1.21 to 1.36.

Groups of 10 can be used instead of groups of 5 by changing all 5's to 10's and changing constants as follows:

Paragraph 5 a - Divide by 10 instead of 5.
Paragraph 10 c - Change .594 to .318.
Paragraph 10 d - Change 2.08 to 1.75 and .254 to .439.
Paragraph 14 - Change .707 to .536; .843 to .637; and 1.21 to .913.

For a given number of observations, the relative precision of results obtained by the use of groups of 4, 5, or 10 under the method outlined is practically the same. However, the smaller groups are to be preferred because of their greater sensitivity to a changing cause system, which is of relatively great importance in manufacture.

Revised 1-15-37.

Leslie E. Simon,
Capt., Ord. Dept.

The following change is added to "Instructions for Control of Quality of Product thru Percentage Inspection".

Interpolate in Paragraph 16.

Groups of six can be used just as readily as groups of five by changing all 5's to 6's and changing constants as follows:

Paragraph 5 a - Divide by 6 instead of 5.
Paragraph 10 c - Change .594 to .498.
Paragraph 10 d - Change 2.08 to 1.97 and .254 to .308.
Paragraph 14 - Change .707 to .649; .843 to .773; and 1.21 to 1.11.

17. Although groups of 4, 5, 6, or 10 are to be preferred, groups of 2 or 3 can be used.

Groups of 2 can be used instead of groups of 5 by changing all 5's to 2's and changing constants as follows:

Paragraph 5 a - Divide by 2 instead of 5.
Paragraph 10 c - Change .594 to 1.94.
Paragraph 10 d - Change 2.08 to 3.52 and .254 to 0.
Paragraph 14 - Change .707 to 1.46; .843 to 1.74; and 1.21 to 2.49.

Groups of 3 can be used instead of groups of 5 by changing all 5's to 3's and changing constants as follows:

Paragraph 5 a - Divide by 3 instead of 5.
Paragraph 10 c - Change .594 to 1.05. 1.58
Paragraph 10 d - Change 2.08 to ~~2.62~~ and .254 to ~~.077~~. 0.10
Paragraph 14 - Change .707 to .972; .843 to 1.16; and 1.21 to 1.66.

18. In some cases where averages ($\bar{X}$'s) are plotted, it becomes necessary to obtain the limits for the chart for averages from the chart itself, instead of from the average range $\bar{W}_t$. This procedure can be accomplished as follows: Ignore the chart for ranges, if one exists. Consider the first 5 plotted averages as observations. Subtract the least from the greatest. Call this value $W_{t_{\bar{X}}}$ (W sub t bar X).

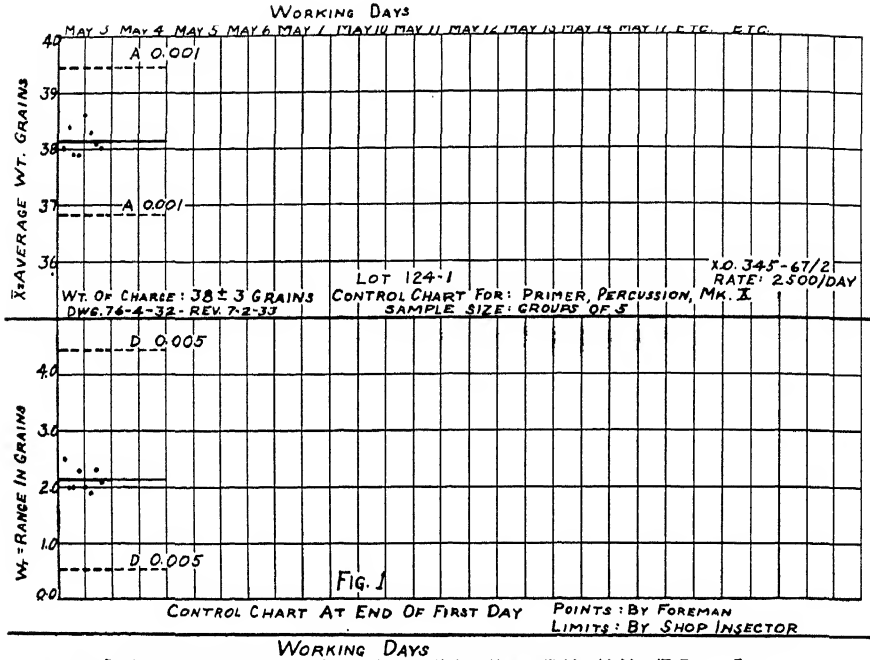
Consider the next 5 plotted averages as observations: subtract the least from the greatest, thereby obtaining another $W_{t_{\bar{X}}}$, etc. Upon the completion of this process for all available plotted averages, compute the average of the $W_{t_{\bar{X}}}$'s. Call this value $\bar{W}_{t_{\bar{X}}}$. The average of the averages $\pm 1.33 \bar{W}_{t_{\bar{X}}}$ will give the A.001 limits for the chart for averages. This method should only be used when there is no chart for ranges or upon the advice of the Department Chief.

TABLE-1

NOTE:

THE INDIVIDUAL READINGS FOR EACH GROUP OF 5 ARE TAKEN. FROM THESE READINGS IS COMPUTED THE INFORMATION TO BE ENTERED ON THE CONTROL CHARTS. THE METHOD OF COMPUTATION IS INDICATED ON THE RIGHT AND INCLUDE THE INDIVIDUAL READINGS FOR THE FIRST DAYS SAMPLING INDICATED ON THE ACCOMPANYING CONTROL CHARTS.

★ 39.0 + 38.0 ★ 36.5 − 37.6 38.9 5)190.0 38.0 = 1ST. $\bar{X}$ ★ 39.0 + 36.5 − 2.5 = 1ST. W_T	★ 39.3 + 38.2 38.0 ★ 37.0 − 37.0 5)189.5 37.9 = 4TH. $\bar{X}$ ★ 39.3 + 37.0 − 2.3 = 4TH. W_T	★ 39.3 + 38.8 38.4 37.0 ★ 37.0 − 5)190.5 38.1 = 7TH. $\bar{X}$ ★ 39.3 + 37.0 2.3 = 7TH. W_T
39.0 ★ 39.5 + 38.0 38.0 ★ 37.5 − 5)192.0 38.4 = 2ND. $\bar{X}$ ★ 39.5 + 37.5 − 2.0 = 2ND. W_T	★ 39.5 + 38.0 39.1 39.0 ★ 37.4 − 5)193.0 38.6 = 5TH. $\bar{X}$ ★ 39.5 + 37.4 − 2.1 = 5TH. W_T	38.1 ★ 38.5 + 38.5 38.5 ★ 36.4 − 5)190.0 38.0 = 8TH. $\bar{X}$ ★ 38.5 + 36.4 − 2.1 = 8TH. W_T
37.5 38.0 ★ 39.0 + 38.0 ★ 37.0 − 5)189.5 37.9 = 3RD. $\bar{X}$ ★ 39.0 + 37.0 − 2.0 = 3RD. W_T	★ 38.9 + 38.4 38.7 38.5 ★ 37.0 − 5)191.5 38.3 = 6TH. $\bar{X}$ ★ 38.9 + 37.0 − 1.9 = 6TH. W_T	


WORKING DAYS
MAY 3 MAY 4 MAY 5 MAY 6 MAY 7 MAY 10 MAY 11 MAY 12 MAY 13 MAY 14 MAY 17 ETC. ETC.
A 0.001
A 0.001
D 0.005
D 0.005
$\bar{X}$ = AVERAGE WT. GRAINS
W_r = RANGE IN GRAINS
40
39
38
37
36
4.0
3.0
2.0
1.0
0.0
WT. OF CHARGE: 38 ± 3 GRAINS
DWG. 76-4-32 - REV. 7-2-33
LOT 124-1
CONTROL CHART FOR: PRIMER, PERCUSSION, MK. I
SAMPLE SIZE: GROUPS OF 5
X.O. 345-67/2
RATE: 2500/DAY
FIG. 1
CONTROL CHART AT END OF FIRST DAY
POINTS: BY FOREMAN
LIMITS: BY SHOP INSECTOR

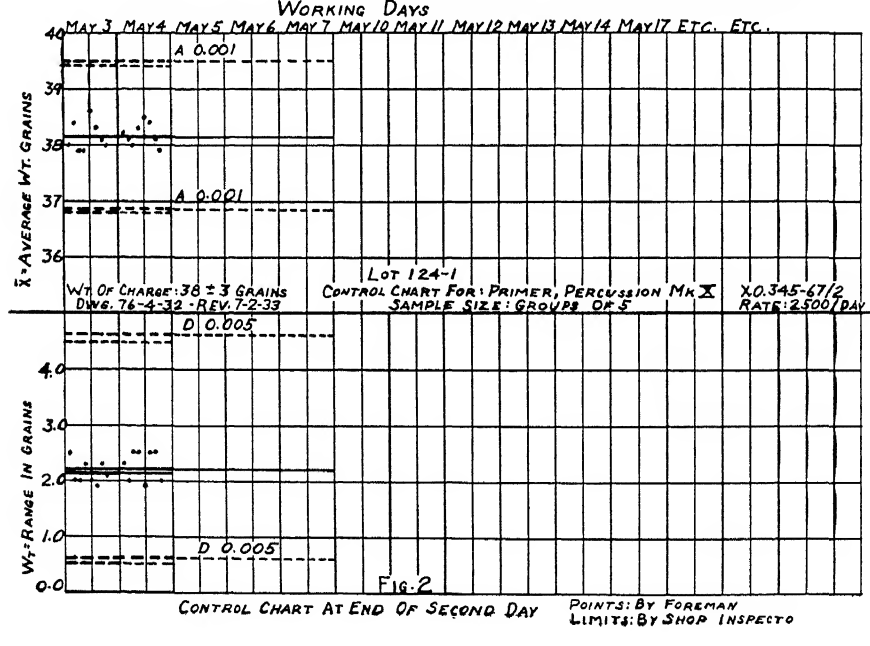

WORKING DAYS
MAY 3 MAY 4 MAY 5 MAY 6 MAY 7 MAY 10 MAY 11 MAY 12 MAY 13 MAY 14 MAY 17 ETC. ETC.
A 0.001
A 0.001
D 0.005
D 0.005
$\bar{X}$ = AVERAGE WT. GRAINS
W_r = RANGE IN GRAINS
40
39
38
37
36
4.0
3.0
2.0
1.0
0.0
WT. OF CHARGE: 38 ± 3 GRAINS
DWG. 76-4-32 - REV. 7-2-33
LOT 124-1
CONTROL CHART FOR: PRIMER, PERCUSSION MK X
SAMPLE SIZE: GROUPS OF 5
X.O. 345-67/2
RATE: 2500/DAY
FIG. 2
CONTROL CHART AT END OF SECOND DAY
POINTS: BY FOREMAN
LIMITS: BY SHOP INSPECTO

ILLUSTRATION OF A SIMPLE QUALITY CONTROL SYSTEM

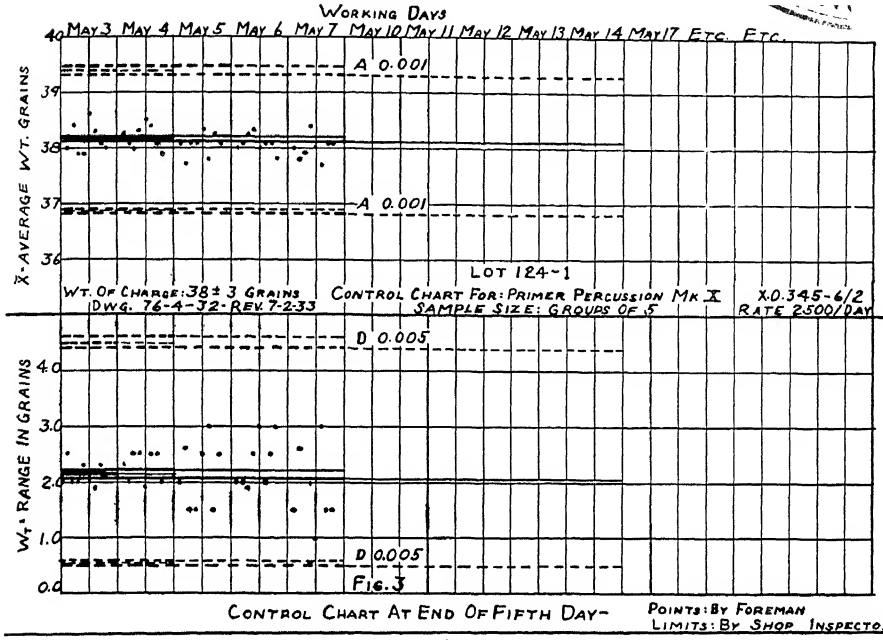


Fig. 3

Control Chart At End Of Fifth Day — Points: By Foreman Limits: By Shop Inspector

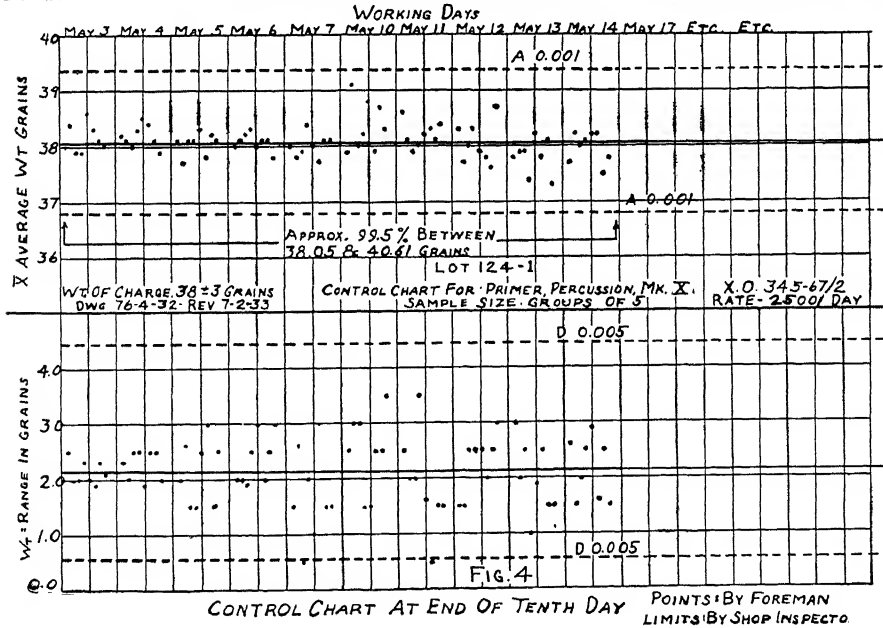


Fig. 4

Control Chart At End Of Tenth Day — Points: By Foreman Limits: By Shop Inspecto

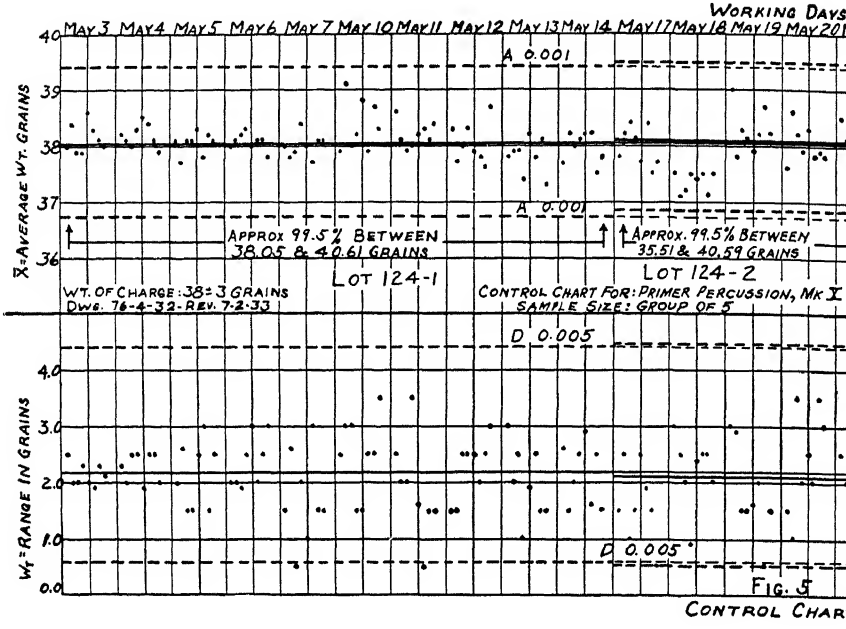
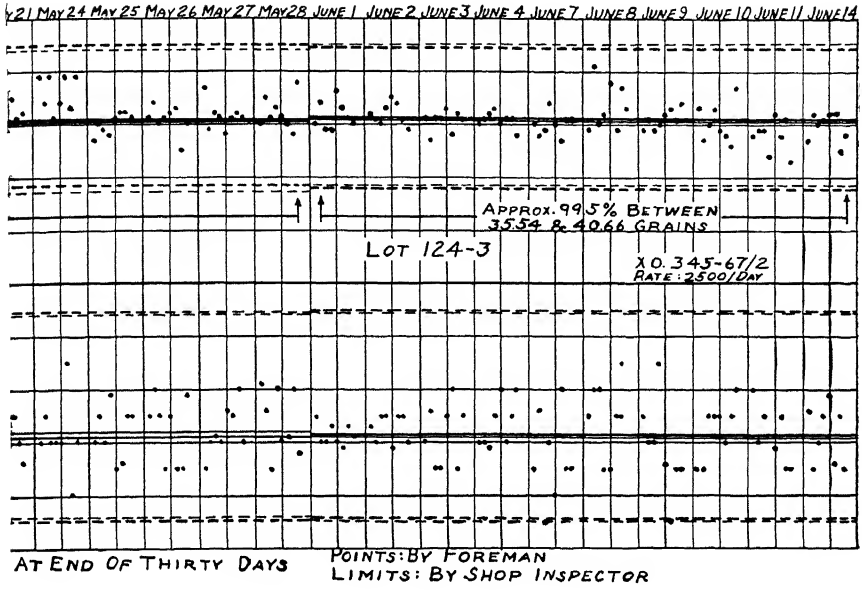


FIG. 5
CONTROL CHART


MAY 21 MAY 24 MAY 25 MAY 26 MAY 27 MAY 28 JUNE 1 JUNE 2 JUNE 3 JUNE 4 JUNE 7 JUNE 8 JUNE 9 JUNE 10 JUNE 11 JUNE 14
APPROX. 99.5% BETWEEN
35.54 & 40.66 GRAINS
LOT 124-3
X 0.345-67/2
RATE: 2500/DAY
AT END OF THIRTY DAYS
POINTS: BY FOREMAN
LIMITS: BY SHOP INSPECTOR

THE STATISTIC RANGE

Published work on the statistic range is attributable for the most part to Pearson and Tippett.[3] They worked from the viewpoint of Pearson-type curves and of obtaining the distribution of range by the method of moments. They tabulated the mean values of range for $n = 2$ to $n = 1000$, the standard deviation of range for a considerable number of values of n and the values of range for a variety of probabilities for $n = 2$ to $n = 100$, thereby laying the foundations for the use of range in control chart technique.

Dr. L. S. Dederick of the Ballistic Research Laboratory, in an unpublished paper, derived the exact distribution of range in convenient form. His interest was from the viewpoint of ballistics, since the statistic range defines the bracket or limits of a salvo for a battery of guns or bombs dropped in mass from an airplane. For the Gaussian distribution his result is:

$$P(R<k) = \frac{n}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{X^2}{2\sigma^2}}\left[P\left(\frac{X}{\sigma}+\frac{k}{\sigma}\right) - P\left(\frac{X}{\sigma}\right)\right]^{n-1} dx\,,$$

where $P(R<k)$ is the probability that the range is less than k and $P(u/\sigma)$ equals the integral of the Gaussian function for the limits zero to any desired value u. This equation yields readily to numerical integration by quadrature. Mr. H. R. Bellinson of the Ballistic Research Laboratory pointed out that solutions other than Gaussian could be obtained from the same distribution function. Dederick tabulated the values of a number of functions of range, and his results, although carried to a greater number of significant figures, agree with those of Tippett and Pearson. In addition, he tabulated the modal and median values.

The tables of factors for ranges. Tables A1 and A2 are an extension of these researches so as to present factors for range in most useful form for control chart work. The control chart limits for averages are merely $\bar{\bar{X}} \pm A\bar{R}$, where $\bar{\bar{X}}$ is the average of the sub-

[3] E. S. Pearson, *Application of Statistical Methods to Industrial Standardisation and Quality Control*, The British Standards Institution, London, 1935.

L. H. C. Tippett, "On the Extreme Individuals and the Range of Samples Taken from a Normal Population," *Biometrika*, Vol. XVII, 1925, pp. 364-387; E. S. Pearson, *Biometrika*, Vol. XVII, pp. 173-194; "Student," *Biometrika*, Vol. XIX, pp. 151-164.

O. L. Davies and E. S. Pearson, "Methods of Estimating from Samples the Population Standard Deviation," *Supplement to the Journal of the Royal Statistical Society*, Vol. I, 1934.

Tables for Statisticians and Biometricians, Part II, pp. CX-CXIX, Tables XXI, XXII, XXIII, and XXIV, Cambridge University Press.

group averages, $\bar{R}$ the average range, and A is given in Table C1. The column headed $A_{3\sigma}$ is supplied for those who prefer to use symmetric multiples of the standard deviation with careful avoidance of any associated probability. The limits, $A_{0.001}$, $A_{0.005}$, are supplied for those who may find slightly different limits justified in practice. It should be noted, however, that the specific numerical probabilities, 0.001 and 0.005, cannot be associated with these limits as actual distributions will not be strictly Gaussian. The author has found the wider limit for averages, $A_{0.001}$, useful in conjunction with the slightly narrower limits for ranges, $D_{0.005}$ and $D_{0.995}$, as, otherwise, indicated lack of control on the chart for averages appeared to cause one to look for trouble unnecessarily with greater frequency than like action was caused by the chart for ranges.

TABLE C1

FACTORS FOR CONTROL LIMITS FOR AVERAGES

(Using Range)

Sample size	$\bar{R}$ in Std. Dev.'s	The limits are $\bar{X} \pm A\bar{R}$						Sample size
		$A_{3\sigma}$	$A_{0.001}$	$A_{0.005}$	$A_{0.01}$	$A_{0.05}$	$A_{0.10}$	
2	1.128	1.880	1.937	1.614	1.458	1.031	0.803	2
3	1.693	1.023	1.054	0.879	0.794	0.561	0.437	3
4	2.059	0.729	0.751	0.626	0.565	0.399	0.311	4
5	2.326	0.577	0.594	0.495	0.447	0.316	0.246	5
6	2.534	0.483	0.498	0.415	0.375	0.265	0.206	6
7	2.704	0.419	0.432	0.360	0.325	0.229	0.179	7
8	2.847	0.373	0.384	0.320	0.289	0.204	0.159	8
9	2.970	0.337	0.347	0.289	0.261	0.185	0.144	9
10	3.078	0.308	0.318	0.265	0.239	0.169	0.132	10
11	3.173	0.285	0.294	0.245	0.221	0.156	0.122	11
12	3.258	0.266	0.274	0.228	0.206	0.146	0.114	12
13	3.336	0.249	0.257	0.214	0.193	0.137	0.107	13
14	3.407	0.235	0.242	0.202	0.183	0.129	0.101	14
15	3.472	0.223	0.230	0.192	0.173	0.122	0.095	15

The additional factors $A_{0.01}$, $A_{0.05}$, and $A_{0.1}$ are supplied for use in inspection techniques. In general, the 5% level is quite useful in detecting significance.

From the column $\bar{R}$ in standard deviations one can readily compute the limits within which approximately any percentage of the product should lie such as those given in paragraph 14 of the illustration. That is, the observed average range divided by $\bar{R}$ for the subgroup size yields an estimate of the universe standard deviation. The approximate percentage of all individuals expected between symmetric limits of $t\sigma$ can be taken from Fig. 10·4. For example, using

TABLE C2

FACTORS FOR CONTROL LIMITS FOR RANGES

Sample size	*Lower Limits* The limit is $\bar{R}D$					*Upper Limits* The limit is $\bar{R}D$					Sample size
	$D_{-3\sigma}$	$D_{0.005}$	$D_{0.01}$	$D_{0.05}$	$D_{0.10}$	$D_{+3\sigma}$	$D_{0.995}$	$D_{0.99}$	$D_{0.95}$	$D_{0.90}$	
2	0	0.009	0.018	0.080	0.160	3.268	3.518	3.226	2.455	2.065	2
3	0	0.100	0.135	0.266	0.372	2.574	2.576	2.422	1.973	1.725	3
4	0	0.185	0.228	0.374	0.476	2.282	2.259	2.128	1.773	1.583	4
5	0	0.254	0.300	0.447	0.542	2.114	2.085	1.973	1.664	1.500	5
6	0	0.308	0.354	0.497	0.588	2.004	1.973	1.870	1.594	1.448	6
7	0.076	0.351	0.395	0.533	0.620	1.924	1.897	1.801	1.544	1.412	7
8	0.136	0.386	0.427	0.562	0.643	1.864	1.837	1.749	1.507	1.384	8
9	0.184	0.415	0.455	0.586	0.663	1.816	1.791	1.707	1.478	1.361	9
10	0.223	0.441	0.480	0.605	0.679	1.777	1.755	1.673	1.455	1.342	10
11	0.255	0.463	0.501	0.621	0.693	1.744	1.724	1.645	1.435	1.327	11
12	0.284	0.482	0.519	0.635	0.706	1.717	1.698	1.621	1.418	1.313	12
13	0.308	0.498	0.534	0.647	0.716	1.692	1.676	1.601	1.403	1.304	13
14	0.329	0.511	0.547	0.658	0.725	1.671	1.656	1.582	1.391	1.294	14
15	0.348	0.524	0.559	0.668	0.732	1.652	1.639	1.567	1.380	1.288	15

subgroups of 5, estimated $\sigma' = \bar{R}/2.326$. From Fig. 10·4, $\pm 2.82\sigma$ includes $99\frac{1}{2}\%$ of the universe (0.25% cut off on each end). Therefore, the average observed range $\pm \bar{R}$ times 2.82/2.326 or $\bar{\bar{X}} \pm 1.21\bar{R}$ includes approximately $99\frac{1}{2}\%$ of all articles.

The control limits for ranges are obtained without the necessity for adding or subtracting. They are merely $\bar{R}D$, where D should be selected in a symmetric manner with respect to approximate probability; i.e., $D_{0.005}$ should go with $D_{0.995}$, etc. The columns $D_{-3\sigma}$ and $D_{+3\sigma}$ are useful in connection with limit inequalities such as the

Tchebycheff theorem. An objection to their use is that, for sample sizes best adapted to range (subgroups of the order of 5), these limits prevent one from detecting an assignable cause of variation which is in the interest of improved uniformity. Such a cause may prove a direct lead to improvement of process. The limits $D_{0.005}$ and $D_{0.995}$ have been found to work well in practice. In control charts on artillery primers covering millions of primers, the number of points outside these limits was in very close agreement with the indicated probability. The charts shown in the sample system are an extract from actual experience with nomenclature and specific technical data altered. The other limits of D are given for the same reasons as the variety of limits for A. A study of the nature of the distribution of range indicates additional reasons (those of efficiency have already been discussed) for preferring subgroups of the order of 5. As the subgroup size increases, the distribution becomes increasingly leptokurtic (peaked), and the tails of the distribution approach the abscissa more slowly. Consequently, deviations much in excess of 3 standard deviations can legitimately occur. The use of range for subgroups in excess of size 10 is not recommended.

APPENDIX D

SPECIFICATIONS AND STANDARDS OF QUALITY[1]

Introduction. Chapter I gave a preview of the services which the statistical method can give the writer or user (either purchaser or vendor) of specifications. These broad aspects of the technique having been shown, it appeared better to defer a detailed discussion of the contribution of statistical methods to specifications, until a minimum of essential statistical methods and basic concepts of frequency distributions were covered, lest frequent side excursions to explain techniques should render the progress of the discussion of specifications tedious.

One generally regards specification as synonymous with inspection specification, i.e., a statement of requirements which the product must meet in order to pass the acceptance test. It is true that the inspection specification may be a necessary and sufficient basis for a "meeting of the minds" between purchaser and vendor and may be all that need be published. However, the attainment of a non-controversial relationship does not mean that the purchaser will secure maximum satisfaction of his economic wants. A good specification has a background which may far exceed the obvious import of the wording of the inspection specification. It requires for its logical composition:

1. A standard of quality.
2. A design specification.
3. An inspection specification.

In this Appendix it is desired to give emphasis to three points: (*a*) the need for the concept of a design specification as separate from the inspection specification; (*b*) the need for a standard of quality; and (*c*) the usefulness of the statistical method in the

[1] Based in part on an address entitled, "Contribution of Statistics to the Development and Use of Purchasing Specifications and Standards of Quality," given by the author at the Symposium on Statistical Methods in Engineering, the University of Pennsylvania Bicentennial Conference, 1940.

logical and scientific promotion of efficiency and orderliness in speaking of and in procuring what one wants. The role of the statistical method in this program can be brought into sharper focus by occasional contrast with the procedure which usually exists in its absence.

The standard of quality may be regarded as providing little more than a datum point from which quality can be reckoned. A clear definition of the current concept of at least one quality level (and preferably the acceptedly satisfactory quality level) is essential to the specification of any definite quality level. It follows, therefore, that the standard of quality is a live and changing thing, not fixed by will or edict, but inherent in a state of current circumstances, just as the manufacturer's control limits (see Chapters IV and VII) were not subject to will, but inherent in the established process. One does not create standards of quality; one discovers and defines them. Shewhart shows that the standard of quality has its origin in: natural law, authority, specification, custom, and precedent.

The design specification is, in brief, a statement of what is wanted. In a hasty consideration, one might confuse this with the inspection specification, from which it is separate and distinct. A producer cannot contract to give a purchaser what he wants; he can only contract to give a purchaser a product which meets certain defined tests. Without the concept of a design specification, however, it is quite impossible to define the inspection requirements most intelligently. The design specification states a quality goal. To state this goal it must enumerate the quality characteristics to be considered, must specify the operational means by which they are to be perceived and the technique of evaluating results. The technique of evaluating results is essentially statistical and involves not only tolerance limits but also statistical measures such as sample size, number of measurements, degree of objective probability, $\bar{X}$, c, etc.

The inspection specification, on the other hand, merely *states the quantity and kind of evidence* which will be accepted as satisfactory proof that the product will meet the design specification goal. Statements which are incapable of operational verification, such as "The materials and workmanship shall be of the best quality," have no place in this matter-of-fact statement of a contractual relation. Note that the inspection specification is predictive in character, and its prediction is that the product will (in the future, i.e., in service) meet the design. It too involves the quality characteristics, specified

operations, and technique of evaluating results. However, neither the operations, the limits, nor statistics are necessarily the same in the two specifications. Perhaps a brief example will clarify these relationships.

Suppose that one wishes to consider the instantaneous point detonating fuze Mk. 0, from the standpoint of percentage of failures to function. An investigation is conducted of the test records of this product and its recorded service use, and of like records of comparable fuzes and fuzes which perform approximately the same function. Suppose that, whereas, in general, failures are due to the fuzes, some failures are due to the nature of the soil on which they impact, so that, as the fuze approaches perfection, an increase in its reliability becomes unimportant and undetectable owing to the operation of external causes of failure. As a result of the investigation it appears that fuzes which give 96% functioning are considered satisfactory, adequate, dependable, and economical by those concerned with the use of the fuze. Fraction effective 0.96 (with its associated operational verification and technique of evaluation) may then be the current standard of quality for this fuze.

Now, suppose that records show that this type of fuze deteriorates in storage at the rate of 0.1% failures the first year, 0.2% failures the second year, 0.4% failures the third year, etc., so that the cost of increased perfection must be balanced against the cost of ultimate replacement (and other incidental expenses). As a result of this investigation it is found that it appears most economical to manufacture fuzes such that when fired with the ammunition components and in the gun for which intended, for impact on normal soil, they give a fraction effective of 0.98. This (with certain other requirements) is the design specification.

The inspection specification must now state the quantity and kind of evidence which will be accepted as proof that a product presented will meet the design specification. This function is distinctly predictive in nature, and in sampling procedure the assurance that a lot of product will meet the design must always be less than certainty. Consequently, an inspection system designed to give some predetermined assurance must be devised. An example was given in the latter part of Chapter I. If the product meets this requirement, it must be accepted, even though it is known that some lots may meet the requirements of the inspection specification, and when consumed, may fall short of the design specification.

In the actual inspection specification the operational verification of meeting the requirement might consist of fuzes fired by initiating the fuze statically and observing its effect on a lead cylinder, instead of firing in a gun on normal soil. In this event various alterations would be made in the requirements to allow for the relationship between the two concepts of performance.

Faults in specifications. The non-statistical type of specification has two faults. First, it attempts to judge each lot on its own merits on a small-sample basis, which, in the light of Chapters II and V, is obviously not an efficient basis for discrimination between good and bad lots even though the lots are homogeneous. The second and corollary fault is that the specification fails to use hindsight as an intelligent guide to foresight; i.e., it fails to make use of available prior knowledge in addition to that supplied by the sample itself (see Chapter III).

The quite non-technical discussion of Chapter I showed the futility of the grabbag type of specification. The organization that attempts to avoid a careful and scientific study of the problem by writing what appears to be a few good, stiff requirements and hoping for the best, is likely not only to fail to get what it wants, and not know what it has, after it has gotten it, but actually to put a penalty on good production. Without a scientific consideration of the problem (and the problem of dealing with a variable product is fundamentally statistical), it was shown that even the brutal extreme of requiring no defectives in a random sample of n only subjected the good, but not quite perfect, producer more mercilessly to the rigors of an unjust chance, and only slightly lessened the extent to which the poor producer could profit from its vagaries.

The role of statistics in specifications and standards of quality. It has been shown in Chapters III, IV, VI, and VII that the technique supplied by the statistical method for checking the validity of accumulated prior knowledge and of combining it with that supplied by the sample is a common-sense rather than a technical procedure. It requires for its operation only three ideas:

a. The concept of statistical uniformity.
b. The operation of an elementary statistical technique.
c. Recognition of a simple empirical law.

Irrespective of homogeneity, one has almost certain (not just probable) knowledge of the product actually sampled. However,

any inference drawn about the remainder of the lot inheres in some relationship, known or unknown, between sample and lot. Unless a product is homogeneous in the sense of being statistically uniform, inferences regarding the lot, based on the observation of samples taken from that lot, must be severely circumscribed, since the reliability of sample quality as a witness of lot quality is unknown.

On the other hand, if (*a*) statistical uniformity exists, and if one knows (*b*) the general level of a quality characteristic and (*c*) the measure of its variation, then elementary statistical theory can predict the limits within which practically all samples of any size, n, should lie. One can obtain the knowledge of the general level of a quality characteristic and the measure of its variation from the accumulation of a large number of small samples. One can then predict the statistical limits of sampling fluctuation for practically all samples of size n. If a sample falls outside these limits, it appears that a state of statistical uniformity does not exist. If, however, no samples fall outside the limits, there is at least no reason for not believing the product to be statistically uniform. This inference is dependent upon the following empirical law.

Extensive experience has shown that, once a state of statistical uniformity is attained in a manufacturing process, the state of quality is inherent in the process itself and cannot be changed without changing the process.

With these 3 ideas, viz., statistical uniformity, a statistical technique, and an empirical law, and with these 3 pieces of knowledge, viz., a state of statistical uniformity, the quality level, and the measure of sampling fluctuations, the small sample becomes a significant index of lot quality. The steps to significance are two. First, the state of statistical uniformity, level of quality, and measure of variation are inferred from the accumulation of samples. (This is the available prior knowledge.) If statistical uniformity cannot be inferred from the data, then efficient discrimination between lots of good and bad quality by means of small samples is impossible. Second, if the quality is satisfactory, and if a subsequent sample falls within limits, there is no reason for believing that the satisfactory quality level has changed, but if it is not within limits, then the quality level appears to have changed from the known satisfactory level. (This inference comes from the information of the sample only.)

These are essentially the principles already stressed in discussing inspection and process control in manufacture. **Concisely, then, one may say that, whereas one cannot judge, except most approximately, the quality of the lot from the small sample, one can judge by statistical methods whether or not there is any reason for believing that the lot is of a different quality level from its predecessors; and, knowing the quality level of the predecessors from accumulated data, one can then infer the quality of the lot with considerable assurance. In this way, hindsight is used as an intelligent guide to foresight.**

Contribution of statistics in the evolution of a standard of quality. In the light of these observations, let us consider the role of statistics in the development of a purchasing specification. One is in an ill position to write a specification unless one knows what one wants in the first place. Even if one knows what one wants, it does little good to specify it, unless there is at least some assurance that it can be met. Hence, that which is ideally desirable must be tempered by that which is practically obtainable. The only way to ascertain what is practically obtainable is through examination of records.[2]

From a long run of records one can readily determine the fraction defective, the average, the standard deviation, or other measures of the principal quality characteristics of the product which should be expected under good manufacturing practice. Without statistical methods a fair consideration of such mass data would be almost impossible. Furthermore, the statistical method is essential as a means of inquiry into the homogeneity of the product, for, without a state of statistical uniformity (as has just been pointed out), one is very limited in drawing conclusions regarding the lot based on the observation of samples.[3] If a homogeneous product of satisfactory quality level can be economically produced under good manufacturing practice, then one is in a position to describe an economic standard of quality. The statistical methods used in connection with the standard of quality throw much light on what kind of a specification one can write; and the standard of quality gives a known starting

[2] When the product is new, resort must be made to the process described under "The Judgment of Statistical Control," Chapter I, pp. 43-46, and Chapter II, pp. 50-79, *Statistical Method*, W. A. Shewhart, Graduate School, Department of Agriculture, Washington, D. C., 1939.

[3] Chapter X, pp. 121-144 et seq., *Economic Control of Quality of Manufactured Product*, W. A. Shewhart, D. Van Nostrand Co., New York, 1931.

point for quality measurements. Hence, though not a part of the specification, it is surely an almost essential adjunct.[4]

Contribution of statistics to the design specification. It is impossible to specify completely a definite series of operations which will certainly detect that a very simple product is or is not of standard quality. Some of the quality characteristics are not capable of definite measurement, and after all, there is a limit to detail. Hence, the specification of the intended quality must be limited to some chosen or principal quality characteristics of the product. This is termed the design specification, as it describes the design which it is intended that the inspection specification shall obtain. Statistics makes an important contribution to the design specification, as the intended or aimed-at values of the quality characteristics of the design and their tolerance limits can be most concisely described by such statistical measures as the average and standard deviation.

It is notable that the statistics of the design specification involve operations. They are operations of measurement such as the measurement of physical quantities: mass, length, etc., or measurements of phenomena, time, velocity, etc. Thus, the design specification expresses quality in terms of something concrete, experienceable; i.e., it states that, if one does so-and-so, thus-and-so shall follow. For example, if one takes n measurements in a certain way on the focal length of a lens, the average of the n measurements shall be within $\bar{X}' \pm E$. **Thus, quality is expressed in the only terms in which it can be clearly conceived, viz., as a succession of perceivable quality phenomena associated with a previously specified set of operations.**[5] Thus the specified quality of a thing should be and is capable of operational verification; and attempts to express quality otherwise by such phrases as "The materials and workmanship shall be of the best," which do not depict the possibility of experiencing a definite and previously conceived succession of associated impacts on one or more of the five senses, are empty gestures. If this is more

[4] Attention is invited to page 13 of Chapter I. Under the subhead, "The two specifications which must always exist," it was stated that careful study showed that the product could be economically manufactured with a uniformity such that $\bar{R}_5 \not> 0.72\% \bar{X}$. This was the design specification. Actual analysis of every lot of that type of product which had ever been tested showed that $0.60\% \bar{X}$ could be expected to be met. This was the *standard of quality*. The specification purposely aimed "below the standard," because it was engineeringly allowable and would facilitate production in a time of national emergency.

[5] See Chapter V, C. I. Lewis, *Mind and the World-Order*, Charles Scribner's Sons, New York, 1929.

like philosophy than statistics, so much the more reason for regarding statistics as common sense reduced to figures.

Contribution of statistics to the inspection specification. The contribution of statistics to the inspection specification is perhaps more marked. In the inspection, which is generally on a percentage basis, it is obviously impossible to determine with certainty that all the product will meet the ideal of the design. Here again it is necessary to compromise with practicality and frankly admit that the logical mission of this specification is a mere statement of the quantity and kind of evidence which will be accepted as a satisfactory indication that the product will meet the standard of quality. It describes no indefinite ideals but confines itself to the delineation of quality characteristics to be experienced, a sequence of operations to be performed upon the product, the interpretation of the results of such operations, and a definition of the limits within which such interpreted results must lie in order for the product to be acceptable. The objectivity of the statistical method makes it ideally adapted to the role of drastically limiting the need for interpretations of requirements. The necessity of interpretations and vague estimates usually opens the field for differences of opinion, whereas the statistical method swiftly produces the estimate of the measure under consideration in a manner superior to unaided judgment and without the bias always associated with personal consideration. This is a great step toward placing both purchaser and vendor on solid ground of mutual trust and understanding.

It is notable that the inspection specification is distinctly predictive in character. By carrying out the dictates of the inspection specification, one acquires a piece of evidence of quality. Based on this evidence of quality (which after the moment of acquisition is in the past), one, at the present time, renders the prediction that the product will, in the future, meet the goal of the design specification; i.e., that certain operations will result in experiencing certain definite successions of phenomena, thereby making the predicted quality capable of operational verification. Any prediction rendered in this uncertain world carries an assurance of less than certainty. It is therefore highly desirable that the inspection specification be written with some degree of objective probability in view, so that one can form some estimate of the degree of belief one may have in the prediction that the product will meet the design. The probability associ-

ated with the symmetric range, $\pm$ 3 standard deviations, is very useful in this regard.

Furthermore, a probability prediction requires for its practical effectiveness an assurance that the product does not appear to be statistically non-uniform. It is therefore desirable that this part of the specification include a test of the control chart character, in order that non-homogeneous lots can be detected and invalid predictions avoided. This step leads to the profitable utilization of existing knowledge (past performance or hindsight) as an intelligent guide to foresight, as was brought out in Chapters III and VI.

In the light of the statistical technique the probability of accepting poorer than standard quality can be consciously minimized to any degree which appears to be economically desirable. One can do much to reduce the chance rejection of good lots and the chance acceptance of bad lots. Without an exhaustive investigation into the reasons for each step, which would be somewhat tedious, let us see one way (a number of others could be suggested) that these principles can be employed. An example of general character, analogous to that given in Chapter I, will serve as an illustration and guide.

A Statistically Sound Inspection Specification. 1. From the first lot of a series, k test samples of n shall be taken as nearly as practicable in order of manufacture, and the fraction defective, q, of each sample computed. The average of all q's shall not exceed the design specification value, q', and no q shall depart from the average by more than 3 times the standard deviation of q.

2. From subsequent lots of a series, a test sample of n shall be taken at random and the fraction defective, q, computed. The observed q shall not depart from the average of all q's by more than 3 times the standard deviation of q (based on all accumulated samples), nor shall it cause the average of all q's to exceed the design specification value, q'.

3. In the event of a rejection, the next lot presented shall be considered a first lot of a new series.

Detailed steps in the specifications and provisions for retests have been omitted in the interest of focusing attention on statistical measures.

Part 1, which involves a relatively large sample, has two objectives. First, to subdivide the lot into subgroups with respect to time, in order that undue variation in the lot from part to part can be detected. Second, to provide a large sample from which quality

can be estimated with a high degree of precision, thereby providing effective discrimination between the manufacturer who has an established process in which the inherent quality level is good, i.e., of specification quality or better, and the manufacturer who is poor and who would otherwise ride as a parasite on the sampling fluctuations provided by a kindly chance. Part 1 makes relatively efficient discrimination on a small-sample basis subsequently possible. It may be noted that the difference in discrimination produced by the relatively large sample of kn is only one of degree, but it is nevertheless an important difference.

Part 2 is a step designed merely to detect a change in quality. It should be noted that it is necessary to take adverse action if there appears to be significant change in quality, even though the sample which is indicative of the change might be acceptable if the average quality level inherent in that manufacturer's process had happened to be poorer but still within specifications. This procedure is necessary since such an instance indicates either a change in average quality or a loss of statistical uniformity and hence results in an unknown situation, and one can no longer validly use the accumulated inspection data in predicting from sample to lot.

Part 3 is, of course, an avenue (though a somewhat painful one) for the manufacturer to reestablish himself.

These measures provide for an adjustment in acceptance criteria whereby the probability of rejection of a product of specification quality can be made very remote without prohibitive sampling. However, in attaining this end, the door has not been opened to the acceptance of poor quality, for: (1) the manufacturer has to qualify first on a large-sample basis; and (2) if he should relax his quality after qualification, he might succeed in passing one or two poor lots, but the probability of passing several poor lots becomes increasingly remote, and once caught, he has to start all over with a new first lot of a series. The extensive operation of these principles at the Ballistic Research Laboratory on accepted stocks of ammunition which were produced under non-statistical specifications not only shows in bold relief the advantages of their employment, but also shows the great gain that would be made if the influx of poor quality were checked at the time of its occurrence.

Check List of Steps in the Determination of a Standard of Quality

1. Collection of data.
   - *a.* Designation of type of product or types of allied products to be investigated.
   - *b.* Designation of size, kind, etc.
   - *c.* Designation of minimum sample size from each source, if more than one source is investigated.
     - (1) Designation of method of dividing samples into subgroups.
       - (i) Arbitrary subgroups.
       - (ii) Subgroups by shipments, lots, consignments, etc.
2. Analysis of data.
   - *a.* Selection of the measure of variation.
     - (1) Standard deviation.
     - (2) Variance.
     - (3) Range, etc.
   - *b.* Method of computation of the measure of dispersion.
     - (1) Large samples.
       - (i) Subgroups of equal size.
       - (ii) Subgroups of unequal size.
     - (2) Small samples.
       - (i) Subgroups of equal size.
       - (ii) Subgroups of unequal size.
   - *c.* Computation of the average.
     - (1) Methods of weighting.
3. Designation of limits for statistical control.
   - *a.* Choice of limits.
   - *b.* Method of computing limits.
4. Action to be taken with regard to data which show lack of control.
5. Basis of recomputing data after elimination of that which,
   - *a.* Is to be rejected for statistical reasons.
   - *b.* Is to be rejected for engineering reasons.
   - *c.* Is to be rejected because it comes from a product not judged to be satisfactory and economic.
6. Brief statement of the standard of quality, preferably in terms of the average and standard deviation of its principal quality characteristics.

Check List of Steps in a Design Specification

1. A study of the standard of quality.
2. A study of the quality required of the product in connection with its intended use.
3. Selection of an economic quality considering increased cost of high quality and decreased utility of low quality.
   - *a.* By mathematical means.
   - *b.* By means of judgment.

4. Expression of quality.
   a. Selection of the quality characteristics to be specified.
   b. Specification of operation of measuring or otherwise experiencing the evidences of the quality characteristic.
   c. Specification of the technique of verification.
      (1) Statistics to be considered.
      (2) Limits of statistics.
   d. Selection of statistics of the specified quality, characteristics: fraction defective, $\overline{X}$, σ, etc.
   e. Method of weighting, combining, and evaluating the aggregate of the quality characteristics.

Check List of Steps in an Inspection Specification

1. Designation of the operational technique of inspection or test.
2. Specification of the quality characteristics to be experienced.
3. Selection of degree of objective probability that the product will meet the design specification.
4. Construction of "operating characteristic of the specification."
5. Measurements to be taken of each quality characteristic of a single article.
   a. Number of measurements.
   b. How taken.
   c. By whom.
6. Number of articles to be measured or inspected.
7. Statistics to be considered.
8. Method of computation of statistics.
9. Method of evaluating results.
   a. Consideration to be given existing knowledge of associated and previously inspected products.
   b. Tests of statistical control.
      (1) Limits within which test results must lie.
10. Conditions under which retests will be permitted.
11. An entire new inspection specification for the retest.
12. Calculation of the change in objective probability, produced by the retest provision.
13. Recast of specification if probability is not satisfactory.

LITERATURE CITED

CAMPBELL, G. A., "Probability Curves Showing Poisson's Exponential Summation, *Bell System Technical Journal*, January, 1923.

CAMP, B. H., "A New Generalization of Tchebycheff's Statistical Inequality," *Bulletin of the American Mathematical Society*, Vol. 28, 1922.

COGGINS, P. P., "Some General Results of Elementary Sampling Theory for Engineering Use," *Bell System Technical Journal*, January, 1928.

COOLIDGE, J. L., *An Introduction to Mathematical Probability*, Oxford Press, London, 1925.

CRANZ, Co., and K. BECKER, *Handbook of Ballistics*, H. M. S. Stationery Office, London, 1921.

CROXTON and COWDEN, *Applied General Statistics*, Prentice-Hall, New York, 1939.

DAVIES, O. L., and E. S. PEARSON, "Methods of Estimating from Samples the Population Standard Deviation," *Supplement to the Journal of the Royal Statistical Society* (Industrial and Agricultural Research Section), Vol. I, No. 1, London, 1934.

DEMING, W. EDWARDS, and RAYMOND T. BIRGE, *On the Statistical Theory of Errors*, Graduate School, Department of Agriculture, Washington, D. C.

DODGE, H. F., "A Method of Rating Manufactured Product," *Bell System Technical Journal*, April, 1928.

FISHER, R. A., *Statistical Methods for Research Workers*, Oliver and Boyd, London, 1936.

FRY, T. C., *Probability and Its Engineering Uses*, D. Van Nostrand Co., New York, 1928.

HELMERT, F. R., *Astronomische Nachrichten* 88, No. 2096, 122, (1876.)

KEYNES, JOHN MAYNARD, *A Treatise on Probability*, Macmillan & Co., London, 1921.

LAPLACE, *Théorie analytique des probabilités*, Gauthier-Villars, Paris, 1886.

LEWIS, C. I., *Mind and the World Order*, Charles Scribner's Sons, New York, 1929.

MCEWEN, G. F., *Methods of Estimating the Significance of Differences in or Probabilities of Fluctuations Due to Random Sampling*, University of California, Berkeley, Calif., 1929.

MEIDELL, M. B., "Sur un problème du calcul des probabilités et statistiques mathématiques," *Comtes Rendus*, Vol. 175, 1922.

MOLINA, E. C., "Bayes' Theorem: An Expository Presentation," *Bell System Technical Journal*, January, 1924.

MOLINA, E. C., and R. I. WILKINSON, "Frequency Distribution of the Unknown Mean of a Sampled Universe," *Bell System Technical Journal*, Vol. 8, October, 1929.

NEYMAN, J., "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Transactions Royal Society*, Vol. A, pp. 289-337, London, February 16, 1933.

1933 A.S.T.M. Manual, *Presentation of Data*, Second Printing, March 1, 1933, A.S.T.M., Philadelphia, Pa.

PEARSON, E. S., *The Application of Statistical Methods to Industrial Standardisation and Quality Control*, British Standards Institution, London, 1935.

Biometrika, Vol. XVII, pp. 173-194.

PEARSON, KARL, *Tables of the Incomplete Beta-Function Ratio*, The Biometrika Office, University College, London, 1934.

Tables for Statisticians and Biometricians, Part I and Part II, Cambridge University Press, London.

POINCARÉ, H., *Calcul des probabilités*, 2ième éd., Gauthier-Villars, Paris, 1912.

Proceedings of the Industrial Statistics Conference, Pitman Publishing Corp., New York, 1939.

SCHROCK, EDWARD M., "Statistical Analysis of Metallurgical Problems," *Metal Progress*, August, 1940.

SHEWHART, W. A., "Some Aspects of Quality Control," *Mechanical Engineering*, December, 1934.

Statistical Method from the Viewpoint of Quality Control, Graduate School, Department of Agriculture, Washington, D. C., 1939.

Economic Control of Quality of Manufactured Product, D. Van Nostrand Co., New York, 1931.

SIMON, L. E., "Deviations in Product Prove Machine Performance," *Product Engineering*, December, 1936.

Contribution of Statistics to the Development and Use of Purchasing Specifications and Standards of Quality, The University of Pennsylvania, Philadelphia, 1940.

"On the Initiation of Statistical Methods for Quality Control in Industry," *Journal of the American Statistical Association*, March, 1941.

STUDENT, *Biometrika*, Vol. VI, pp. 1-25, 1908; Vol. XIX, pp. 151-164.

TIPPETT, L. H. C., "On the Extreme Individuals and the Range of Samples Taken from a Normal Population," *Biometrika*, Vol. XVII, pp. 364-387, 1925.

USPENSKY, J. V., *Introduction to Mathematical Probability*, McGraw-Hill Book Co., N. Y., 1937.

WHITTAKER, E. T., and G. ROBINSON, *The Calculus of Observations*, Blackie & Sons, London, 1929.

YULE, G. UDNY, and M. G. KENDALL, *An Introduction to the Theory of Statistics*, XI Edition, Chas. Griffin & Co., 1937.

GLOSSARY OF SYMBOLS

A	arbitrary constant.
$A_{0.005}$	a factor for control limits for averages using range.
a'	fraction of a lot.
a	an arbitrary constant. Also used to designate the average number of defectives.
B	arbitrary constant.
b	an arbitrary constant.
C	cost of inspection plus average cost of misgrading.
$\bar{c}'$	the true average number of defects or defectives in a lot or universe.
$\bar{c}$	the average number of defects or defectives, usually the average number in a sample of n.
c	the number of defects or defectives, usually the number of defectives in a sample of n.
$c_{L0.1}$	lower limit of c for the probability 0.1.
$c_{U0.9}$	upper limit of c for the probability 0.9.
$c_{L0.005}$	lower limit of c for the probability 0.005.
$c_{U0.995}$	upper limit of c for the probability 0.995.
c_2	a factor which is a function of n, and expresses the ratio of $\bar{\sigma}$ to σ'.
$D_{0.005}$	a factor for control limits for ranges using range.
E	an arbitrary constant generally used to designate the magnitude of an error.
$I_x(\mathrm{p}, q)$	the incomplete beta-function ratio.
I_Q	the symbol for the incomplete beta-function ratio.
$I_Q(c, n-c+1)$	symbolizes $\int_0^Q Q^{c-1}(1-Q)^{n-c} \Big/ \int_0^1 Q^{c-1}(1-Q)^{n-c}$; see Appendix B.
$I_Q(c, n-c+1)$	the probability of at least c defectives in a random sample of n.
i	arbitrary constant.
J_0	the second moment about the mean.
J_D	the second moment about a point the distance D from the mean.
j	arbitrary constant.
k	arbitrary constant.
$L_1 - L_2$	tolerance limits.
l	arbitrary constant.
M	the cost of misgrading an article.
m	arbitrary constant.
N	the number of articles in a population, lot, or universe.
n	number of articles in a sample.
$P_{t\sigma}$	the probability associated with a symmetric range of $t\sigma$.
$P(c, \infty, a)$	the probability of at least c defectives in an infinite sample when the average number in such sample is a.

$P(c, n, a)$	the probability of at least c defectives in a sample of n when the average number is a.
$P(> c - 1)$	probability of at least c defectives.
P_m	the probability of misgrading a lot of articles.
P_s	a probability level of significance.
P_d	the probability of a difference as great as or greater than that observed.
P	the true lot fraction effective, i.e., the true proportion of defective articles in the lot or universe.
p	the sample fraction effective, i.e., the proportion of defective articles in the sample.
Q	the true lot fraction defective, i.e., the true proportion of defective articles in the lot or universe.
Q_M	the middle-most value of the lot fraction defective often used as the estimated lot fraction defective.
Q_L	a lower limit of estimated lot fraction defective.
Q_U	upper limit of estimated lot fraction defective.
q	the sample fraction defective, i.e., the proportion of defective articles in the sample.
R	arbitrary constant.
$\overline{R}$	average range.
r	the correlation coefficient.
T	the cost of selection and inspection of an article.
t	an arbitrary constant generally used to designate the number of standard deviations.
v	arbitrary constant.
$W(x)$	the a priori probability that Q equals x.
$W(Q_L, Q_U)$	the probability that true Q lies between Q_L and Q_U.
$\overline{X}^2$	the square of the average of X's.
$\overline{X^2}$	the average of squares of X's.
X	a single measure or observation of a quality characteristic; a member of a frequency distribution.
$\overline{X}$	the sample average of two or more X's.
$\overline{X}'$	the true lot or population average.
$\overline{\overline{X}}$	the average of two or more $\overline{X}$'s.
x	the deviation of an X from $\overline{X}$ (exception, used as a variable in Appendix B).
$1/x'$	fraction of a success.
$\bar{x}$	the average of two or more deviations of X from $\overline{X}$.
x/σ	the deviation of an X in standard deviations as a unit.
y	the deviation of a Y from a $\overline{Y}$.
$\overline{Y}^2$	the square of the average of Y's.
$\overline{Y^2}$	the average of squares of Y's.
$\mathrm{B}_x(\mathrm{p}, q)$	the incomplete beta function.
$\mathrm{B}(\mathrm{p}, q)$	the complete beta function.
Δ	the numerator of a fraction resulting from the difference of two ratios.
θ	any statistic.

Σ	summation of a discrete variable.
σ	the observed standard deviation.
σ_{xs}	the standard deviation of the X values about the regression line.
σ_{ys}	the standard deviation of the Y values about the regression line.
σ_r	the standard deviation of the correlation coefficient.
$\sigma(\delta)$	the standard deviation estimated from successive differences.
$\overline{\sigma^2(\delta)}$	the average squared standard deviation estimated from successive differences.
σ'	the true standard deviation of the lot or universe.
σ_σ	the observed standard deviation of the standard deviation.
$\bar{\sigma}$	the average of two or more observed standard deviations.
σ_D	the standard deviation of a difference of two variables.
σ_a	the standard deviation of an accidental error.
δ^2	the mean square successive difference.
$\Gamma(c+1)$	gamma function $(c+1)$: equals $c!$ for positive integers.
σ_T	the standard deviation of the total variation.
σ_s	the standard deviation of a systematic error.
!	factorial; e.g., $n!$ is the product of all the natural numbers from 1 to n inclusive. 0! is 1.
$>$	greater than.
$<$	less than.
$\not>$	not greater than.
$\neq$	not equal.

A bar over a statistic, e.g., $\overline{X}$, is used to indicate the average of that statistic.

A prime on a statistic, e.g., $\overline{X}'$ is used to designate the true but unknown universe value of that statistic.

INDEX

S

T

U

V